

日本国特許庁
JAPAN PATENT OFFICE

J. Yamato et al.

8/21/03

Q76950

1 of 1

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出願年月日

Date of Application:

2002年 8月28日

出願番号

Application Number:

特願2002-249049

[ST.10/C]:

[JP 2002-249049]

出願人

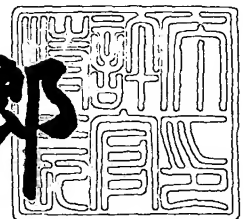
Applicant(s):

日本電気株式会社

2003年 5月20日

特許庁長官
Commissioner,
Japan Patent Office

太田信一郎



出証番号 出証特2003-3037425

【書類名】 特許願

【整理番号】 35001164

【提出日】 平成14年 8月28日

【あて先】 特許庁長官 殿

【国際特許分類】 G06F 13/00

【発明者】

 【住所又は居所】 東京都港区芝五丁目 7 番 1 号 日本電気株式会社内

 【氏名】 大和 純一

【発明者】

 【住所又は居所】 東京都港区芝五丁目 7 番 1 号 日本電気株式会社内

 【氏名】 菊地 芳秀

【発明者】

 【住所又は居所】 東京都港区芝五丁目 7 番 1 号 日本電気株式会社内

 【氏名】 金子 裕治

【特許出願人】

 【識別番号】 000004237

 【氏名又は名称】 日本電気株式会社

【代理人】

 【識別番号】 100103090

 【弁理士】

 【氏名又は名称】 岩壁 冬樹

 【電話番号】 03-3811-3561

【選任した代理人】

 【識別番号】 100114720

 【弁理士】

 【氏名又は名称】 須藤 浩

 【電話番号】 03-3811-3561

【手数料の表示】

 【予納台帳番号】 050496

【納付金額】 21,000円

【その他】 国等の委託研究の成果に係る特許出願（平成13年度新エネルギー・産業技術総合開発機構「基盤技術研究促進事業（民間基盤技術研究支援制度）「大規模・高信頼サーバの研究」」に関する委託研究、産業活力再生特別措置法第30条の適用を受けるもの）

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【包括委任状番号】 0102926

【ブルーフの要否】 要

【書類名】 明細書

【発明の名称】 データ複製システム、中継装置、データ送受信方法およびストレージ内のデータを複製するためのプログラム

【特許請求の範囲】

【請求項 1】 第一のストレージ内のデータを通信ネットワークを介して第二のストレージにミラーリングまたはバックアップするデータ複製システムにおいて、

前記第一のストレージから前記第二のストレージに転送されるデータを中継する中継装置であって、前記第一のストレージが災害によって稼働できない状態になっても、稼働を継続できるとあらかじめ算定された位置に設置された中継装置を備え、

前記第一のストレージは、データ転送の制御を行うデータ転送処理手段を含み、

前記データ転送処理手段は、前記中継装置に対してデータ転送を完了したときに、前記第二のストレージに対してデータ転送を完了したとみなすことを特徴とするデータ複製システム。

【請求項 2】 中継装置は、第一のストレージから受信したコマンドおよびデータを記憶するための不揮発性記憶手段と、データの中継制御を行う中継処理手段とを含み、

前記中継処理手段は、第一のストレージから受信したコマンドおよびデータを前記不揮発性記憶手段に記憶し、記憶されたコマンドおよびデータを任意のタイミングで第二のストレージに送信する

請求項 1 記載のデータ複製システム。

【請求項 3】 複数の中継装置が設けられ、

第一のストレージにおけるデータ転送処理手段は、第一のストレージ内のデータを同時に複数の中継装置に向けて送信する

請求項 1 または請求項 2 に記載のデータ複製システム。

【請求項 4】 第一のストレージから第二のストレージに転送されるデータを中継する中継装置であって、

前記第一のストレージから受信したデータを記憶するための記憶手段と、データの中継制御を行う中継処理手段とを含み、

前記中継処理手段は、前記第一のストレージから受信したデータを記憶手段に記憶し、記憶手段にデータを記憶したときに前記第一のストレージに対する応答を送信し、前記第二のストレージに対して記憶手段に記憶したデータを送信することを特徴とする中継装置。

【請求項 5】 データを送信する送信元からデータを受信する送信先にデータを送信するデータ送受信方法において、

前記送信元では、送信される元データから少なくとも 1 つのエラー訂正のための冗長データを作成し、元データと冗長データとを別々のデータ送信単位で送信する

ことを特徴とするデータ送受信方法。

【請求項 6】 送信先では、元データと冗長データとの集合であるデータ群のすべてについて受信を完了する前に、元データについて部分的にエラー訂正処理を実行できる前記データ群の一部を受信した段階で、エラー訂正処理を実行する請求項 5 記載のデータ送受信方法。

【請求項 7】 送信元では、元データを分割データに分割し、それらの分割データのうちの 1 つまたは複数が消失しても元データを復元可能な冗長データを作成する

請求項 5 または請求項 6 に記載のデータ送受信方法。

【請求項 8】 冗長データとしてパリティデータまたは ECC を用いる請求項 7 記載のデータ送受信方法。

【請求項 9】 冗長データとして送信データの複製データを用いる請求項 5 から請求項 7 のうちのいずれか 1 項に記載のデータ送受信方法。

【請求項 10】 元データと冗長データとを、別々の通信ネットワークに送出する

請求項 5 から請求項 9 のうちのいずれか 1 項に記載のデータ送受信方法。

【請求項 11】 第一のストレージ内のデータを第二のストレージに通信ネットワークを介してミラーリングまたはバックアップするデータ複製システムにお

いて、

前記第一のストレージは、データ転送の制御を行うデータ転送処理手段と、送信される元データから少なくとも1つのエラー訂正のための冗長データを作成する冗長化手段とを含み、

前記データ転送処理手段は、元データと前記冗長化手段が作成した冗長データとを別々のデータ送信単位で送信する

ことを特徴とするデータ複製システム。

【請求項12】 第二のストレージは、第一のストレージから受信した冗長データを用いてエラー訂正処理を行うデータ復元手段と、データ復元手段が復元したデータを記憶媒体に格納する格納処理手段とを含み、

前記データ復元手段は、前記第一のストレージから元データと冗長データとの集合であるデータ群のすべてについて受信を完了する前に、元データについて部分的にエラー訂正処理を実行できる前記データ群の一部を受信した段階で、エラー訂正処理を実行する

請求項11記載のデータ複製システム。

【請求項13】 第一のストレージにおける冗長化手段は、元データを分割データに分割し、それらの分割データのうちの1つまたは複数が消失しても元データを復元可能な冗長データを作成する

請求項11または請求項12に記載のデータ複製システム。

【請求項14】 冗長化手段は、冗長データとしてパリティデータまたはECCを用いる請求項13記載のデータ複製システム。

【請求項15】 冗長化手段は、冗長データとして、元データの複製データを作成する

請求項11から請求項13のうちのいずれか1項に記載のデータ複製システム

【請求項16】 データ転送処理手段は、元データと冗長データとを、別々の通信ネットワークに送出する

請求項11から請求項15のうちのいずれか1項に記載のデータ複製システム

【請求項 1 7】 稼働系のストレージ内のデータを通信ネットワークを介して待機系のストレージにミラーリングするデータ複製システムにおいて、

前記稼働系のストレージは、データの書き込み要求が発生すると、書き込み対象のデータと遅延書き込み要求とを前記待機系のストレージに送信する遅延書き込み要求手段と、そのデータの状態であればアプリケーションがそのまま動作を再開可能な再開可能ポイントであることを知らせるための再開可能ポイント通知を上位装置から受けると遅延書き込み実行要求を前記待機系のストレージに送信する書き込み実行要求手段とを含み、

前記待機系のストレージは、データを一時的に記憶する一時記憶手段と、受信したデータを遅延書き込み要求に応じて前記一時記憶手段に格納するとともに、遅延書き込み実行要求を受信すると前記一時記憶手段に格納されているデータを記憶媒体に格納する格納処理手段とを含む

ことを特徴とするデータ複製システム。

【請求項 1 8】 稼働系のストレージにおける遅延書き込み要求手段と遅延書き込み実行要求手段とは、遅延書き込み要求と遅延書き込み実行要求とを待機系のストレージに非同期に送信し、

前記待機系のストレージにおける格納処理手段は、一の遅延書き込み実行要求を受信すると、一つ前の遅延書き込み実行要求と前記一の遅延書き込み実行要求との間に送信された遅延書き込み要求に対応するデータを全て一時記憶手段に格納し終え、前記一つ前の遅延書き込み実行要求以前に送信されたデータを記憶媒体に格納し終えたときに、一時記憶手段に格納されているデータを記憶媒体に格納する

請求項 1 7 記載のデータ複製システム。

【請求項 1 9】 待機系のストレージにおける格納処理手段は、稼働系に異常が生ずると、一時記憶手段に格納されているデータを破棄する

請求項 1 7 または請求項 1 8 に記載のデータ複製システム。

【請求項 2 0】 稼働系のストレージ内のデータを通信ネットワークを介して待機系のストレージにミラーリングするデータ複製システムにおいて、

前記稼働系のストレージは、データの書き込み要求が発生すると、書き込み対

象のデータと書き込み要求とを前記待機系のストレージに送信する書き込み要求手段と、そのデータの状態であればアプリケーションがそのまま動作を再開可能な再開可能ポイントであることを知らせるための再開可能ポイント通知を上位装置から受けるとスナップショット作成要求を前記待機系のストレージに送信するスナップショット作成要求手段とを含み、

前記待機系のストレージは、書き込み要求を受信したら前記書き込み要求に対応するデータを書き込むべき領域を割り当てて記憶媒体に前記データを格納するとともに、前記記憶媒体へのデータ格納状況を示す格納情報を更新し、スナップショット作成要求を受信したらスナップショットを作成するスナップショット作成手段を含む

ことを特徴とするデータ複製システム。

【請求項 2 1】 書き込み要求手段は、待機系ストレージのスナップショット作成手段がスナップショットを作成している場合には、スナップショット作成手段がスナップショット作成を完了した後に、書き込み要求を待機系のストレージに送信する請求項 2 0 に記載のデータ複製システム。

【請求項 2 2】 稼働系のストレージにおける書き込み要求手段とスナップショット作成要求手段とは、書き込み要求とスナップショット作成要求とを待機系のストレージに非同期に送信し、

前記待機系のストレージにおけるスナップショット作成手段は、各書き込み要求を受信したときに、受信した書き込み要求の直前のスナップショット作成要求に基づくスナップショットの作成を完了するまで、書き込み要求に対応するデータを記憶媒体に格納するのを待ち、

一のスナップショット作成要求を受信すると、一つ前のスナップショット作成要求に基づくスナップショットの作成を完了し、前記一つ前のスナップショット作成要求と前記一のスナップショットとの間に送信された書き込み要求に対応するデータを全て記録媒体に格納し終えたときに、スナップショットを作成する請求項 2 0 に記載のデータ複製システム。

【請求項 2 3】 待機系のストレージにおけるスナップショット作成手段は、稼働系に異常が生ずると、直前のスナップショット作成後にデータが格納された

記憶媒体の領域を未使用状態として解放し、格納情報を直前のスナップショット作成時の状態に戻す請求項 2 0 ないし請求項 2 2 のうちのいずれか 1 項に記載のデータ複製システム。

【請求項 2 4】 稼働系のストレージを使用する稼働系の上位装置は、再開可能ポイントになると前記稼働系のストレージに再開可能ポイント通知を送信する再開可能ポイント通知手段を含み、

待機系のストレージを使用する待機系の上位装置は、前記稼働系の異常を検出すると、異常の発生を待機系のストレージに通知して待機系のストレージの状態を前記再開可能ポイントに対応する状態にするように促し、待機系のストレージの状態が前記再開可能ポイントに対応する状態になったときに処理を再開する

請求項 1 9 または請求項 2 3 に記載のデータ複製システム。

【請求項 2 5】 稼働系のストレージを使用する稼働系の上位装置は、待機系のストレージを使用する待機系の上位装置に、稼働系の処理実行状態を示す実行イメージを転送する実行イメージ転送手段と、

実行イメージ転送手段が実行イメージを転送するタイミングで、稼働系のストレージに再開可能ポイント通知を送信する任意時点再開可能ポイント通知手段とを含み、

待機系のストレージを使用する待機系の上位装置は、前記実行イメージ転送手段が転送する実行イメージを保存する実行イメージ保存手段を含み、

実行イメージ転送手段は、任意の時点における実行イメージを待機系の上位装置に転送し、

待機系のストレージを使用する待機系の上位装置は、前記稼働系の異常を検出すると、異常の発生を待機系のストレージに通知して、待機系のストレージの状態を実行イメージの転送タイミングに対応する状態にするように促し、待機系のストレージの状態が実行イメージの転送タイミングに対応する状態になった場合に実行イメージを用いて処理を再開する

請求項 1 9 または請求項 2 3 に記載のデータ複製システム。

【請求項 2 6】 稼働系の上位装置における実行イメージ転送手段は、前回転送した実行イメージから変更された部分のみを転送する

請求項 2 5 記載のデータ複製システム。

【請求項 2 7】 第一のストレージ内のデータを通信ネットワークおよび中継装置を介して第二のストレージにミラーリングまたはバックアップするデータ複製システムにおける前記中継装置内に設けられているコンピュータに、

前記第一のストレージから受信したデータを中継装置内の記憶媒体に記憶する処理と、

データを中継装置内の記憶媒体に記憶したときに前記第一のストレージに対する応答を送信する処理と、

中継装置内の記憶媒体に記憶したデータを前記第二のストレージに送信する処理とを

実行させるストレージ内のデータを複製するためのプログラム。

【請求項 2 8】 第一のストレージ内のデータを第二のストレージに通信ネットワークを介してミラーリングまたはバックアップするデータ複製システムにおける前記第一のストレージ内に設けられているコンピュータに、

送信される元データから少なくとも 1 つのエラー訂正のための冗長データを作成する処理と、

元データと冗長データとを別々のデータ送信単位で送信する処理とを実行させるストレージ内のデータを複製するためのプログラム。

【請求項 2 9】 稼働系のストレージ内のデータを通信ネットワークを介して待機系のストレージにミラーリングするデータ複製システムにおける前記稼働系のストレージ内に設けられているコンピュータに、

データの書き込み要求が発生すると、書き込み対象のデータと一時記憶装置にデータを格納することを指示する遅延書き込み要求とを前記待機系のストレージに送信する処理と、

そのデータの状態であればアプリケーションがそのまま動作を再開可能なポイントであることを知らせるための再開可能ポイント通知を上位装置から受けると、一時記憶装置に格納されているデータを記憶媒体に格納することを指示する遅延書き込み実行要求を前記待機系のストレージに送信する処理とを実行させるストレージ内のデータを複製するためのプログラム。

【請求項 3 0】 稼働系のストレージ内のデータを通信ネットワークを介して待機系のストレージにミラーリングするデータ複製システムにおける前記稼働系のストレージ内に設けられているコンピュータに、

データの書き込み要求が発生すると、書き込み対象のデータと記憶媒体にデータを書き込むことを指示する書き込み要求とを前記待機系のストレージに送信する処理と、そのデータの状態であればアプリケーションがそのまま動作を再開可能なポイントであることを知らせるための再開可能ポイント通知を上位装置から受けるとスナップショット作成要求を前記待機系のストレージに送信する処理を実行させるストレージ内のデータを複製するためのプログラム。

【発明の詳細な説明】

【 0 0 0 1 】

【発明の属する技術分野】

本発明は、データ複製システム、中継装置、データ送受信方法およびストレージ内のデータを複製するためのプログラムに関する。

【 0 0 0 2 】

【従来の技術】

災害等が発生してもコンピュータシステムの機能を維持できるようにするため、正常系（稼働系）のシステムと待機系のシステムとが設けられたコンピュータシステムが実現されている。例えば、EMC Corporation（イーエムシー コーポレーション）は、正常系のストレージと待機系のストレージとを用いてミラーリングを行うシステムを実現している。このシステムに関する情報は、「http://www.emc2.co.jp/local/ja/JP/products/product_pdfs/srdf/srdf.pdf」という URL で公開されている。

【 0 0 0 3 】

また、特開 2 0 0 0 - 3 0 5 8 5 6 公報には、メインセンター（正常系のシステム）とリモートセンター（待機系のシステム）とでデータの二重化を図るシステムが記載されている。

【 0 0 0 4 】

一般に、正常系のシステムでは、正常系のストレージとそのストレージを使用

する正常系のホストコンピュータ（以下、ホストという。）とが接続される。待機系のシステムも同様である。そして、正常系のストレージと待機系のストレージとが、例えば、専用回線やインターネット等の通信ネットワークを介して接続される。

【0005】

「http://www.emc2.co.jp/local/ja/JP/products/product_pdfs/srdf/srdf.pdf」において公開されているシステムや、特開2000-305856公報に記載されたシステムでは、正常系のシステムから待機系のシステムに直接データを送信している。しかし、正常系のシステムから中継装置にデータを送信し、中継装置から待機系のシステムにデータを送信することもある。一般に、中継装置は、正常系のシステムから受信したデータを待機系のシステムに送信し、待機系のシステムから受信完了の通知を受けたときに、正常系のシステムに待機系へのデータ転送が完了したことを通知している。そして、正常系のシステムは、中継装置にデータを送信した場合、データ転送が完了した旨の通知を中継装置から受信してから次の処理を開始する。

【0006】

また、一般に、通信ネットワークを介してデータを送受信する場合、データを分割してデータ送信単位毎に送受信する。このデータ送信単位は、通信プロトコル毎に異なる。ここでは、パケットをデータ送信単位とする場合を例に説明する。パケット等のデータ送信単位には、送信対象のデータだけでなく、送信過程でのデータの誤り（データ化け）を検出するためのエラー検出コードも含まれる。エラー検出コードとしては、チェックサムデータやCRC（Cyclic Redundancy Check）データ等がある。データを受信した装置は、エラー検出コードによってデータの誤りを検出すると、そのパケットを廃棄する。

【0007】

また、データに誤りが生じていない場合でも、通信ネットワークに輻輳（通信負荷が高い状態）が生じると、パケットは通信ネットワーク上で廃棄される。パケットが廃棄され、送信先からの応答が得られない場合、送信元は再度パケットを送信する。なお、通信ネットワークに輻輳が生じたときにパケットの廃棄を開

始すると、送信元による再送信によってさらに通信負荷が高くなったり、通信ネットワークの利用率が低くなったりすることがある。このような問題を回避するため、近年、通信負荷が所定のしきい値を越えたときに通信ネットワークを構成する機器が任意にパケットを廃棄する方式が採用されている。この方式は、RED (Random Early Detection) と呼ばれている。

【 0 0 0 8 】

また、データを受信した装置がデータの誤りを訂正することができるデータ伝送方式も提案されている。例えば、特開昭 5 7 - 1 3 8 2 3 7 号公報には、送信対象のデータと、そのデータから作成されたパリティビットとを別々に送信し、送信対象のデータとパリティビットとを受信した装置が誤り訂正を行うデータ伝送方式が記載されている。

【 0 0 0 9 】

なお、以下、ミラーリングとバックアップとを以下のように区別するものとする。「ミラーリング」とは、あるストレージに対してホストから書込コマンド (write コマンド) が出力されたことを契機として、そのストレージを含む 2 つ以上のストレージに対して同じデータを書き込むことと定義する。レプリケーションもミラーリングの一種として扱う。また、「バックアップ」とは、ストレージに対する write コマンドを契機とせず、任意のタイミングで、あるストレージの内容を他のストレージに複写することと定義する。

【 0 0 1 0 】

また、コンピュータシステムにおいて、実行中のプログラムを任意の時点で中断させ、後に再開させる技術が実現されている。例えば、S X シリーズという名称で販売されている日本電気株式会社製のスーパーコンピュータでは、任意の時点におけるプロセスの実行状態 (例えば、メモリやレジスタの状態等) を保存して、プログラムを中断・再開するようにしている。この機能は、チェックポイント・リスタート機能と呼ばれる場合もある。

【 0 0 1 1 】

【発明が解決しようとする課題】

正常系のシステムから中継装置を介して待機系のシステムにデータを送信する

場合には、以下のような問題があった。中継装置は、正常系のシステムから受信したデータを待機系のシステムに送信し、待機系のシステムから受信完了の通知を受けたときに、転送が完了したことを正常系のシステムに通知する。そして、正常系のシステムは、転送が完了した旨の通知を中継装置から受信しなければ、次の処理に移行しない。そのため、正常系のシステムが中継装置にデータを送信してから、次の処理を開始できるまでの時間がかかってしまうという問題があった。特に、災害対策として中継装置や待機系のシステムを設ける場合には、正常系のシステムの遠隔地に中継装置を配置し、待機系のシステムをより遠い場所に配置する。その結果、正常系システムと中継装置との間でデータや通知を送受信する時間や、中継装置と待機系システムとの間でデータや通知を送受信する時間がかかってしまい、正常系のシステムにおける次の処理の開始が遅れてしまう。

【0012】

例えば100km離れた地点との通信におけるラウンドトリップ時間（データを送信してから応答を得るまでの往復時間）は、ms（ミリ秒）のオーダーである。ホストは、 μ s（マイクロ秒）のオーダーで各処理を進める。従って、送信に対する応答を得るまでの時間は、処理の遅れの原因となる。

【0013】

また、データの送信過程において、パケット等のデータ送信単位で廃棄されると、送信元となる正常系のストレージはデータを再送信することになる。すると、データ送信完了までに一層時間がかかり、次の処理の開始がさらに遅れてしまう。

【0014】

また、正常系のストレージから待機系のストレージにデータのバックアップ処理を行うときには、データの送信距離が長いことに加えて、送信すべきデータの量が多くなるため、送信時間が一層増大してしまう。

【0015】

また、正常系のシステムに異常が生じたときに、待機系のシステムですぐに処理を開始できるとは限らなかった。例えば、ある処理Xを開始する際には、ストレージに対するデータAおよびデータBの書き込みが完了していなければならな

いとする。正常系のホストがデータ A を書き込む write コマンドをストレージに出力してミラーリングを行うと、待機系のストレージにもデータ A は反映される。その後、正常系のシステムに異常が発生したとする。この場合、待機系のストレージには、データ B が書き込まれていないため、すぐに処理 X を開始できない。待機系のシステムで処理を開始する場合には、待機系のストレージにデータ B を書き込んで処理 X を開始するか、あるいは、データ A を削除して処理 X の前の処理からやり直す必要がある。そのため、待機系システムでの処理の再開に時間がかかってしまっていた。

【 0 0 1 6 】

また、所定のデータ記録状態になっていれば処理を再開できるような機能がアプリケーションプログラムによって実現されていない場合がある。例えば、上記の例においてホストがストレージに A, B を書き込んだならば処理 X から処理を再開できるようにした機能が実現されていない場合がある。以下、ストレージが所定のデータ記録状態になっていれば処理を再開できるような機能を「再開機能」と記す。アプリケーションプログラムによって再開機能が実現されていない場合においても、正常系で異常が生じたならば、待機系で迅速に処理を再開できることが好ましい。

【 0 0 1 7 】

そこで本発明は、中継装置を介して待機系のシステムにデータを送信する場合に、正常系のシステムが早く次の処理を開始できるようにすることを目的とする。また、送信過程でデータが廃棄されたとしても迅速にデータの送受信を完了できるようにすることを目的とする。また、正常系のシステムで異常が発生したときに、すぐに待機系のシステムが処理を進められるようにすることを目的とする。

【 0 0 1 8 】

【課題を解決するための手段】

本発明によるデータ複製システムは、第一のストレージ内のデータを通信ネットワークを介して第二のストレージにミラーリングまたはバックアップするデータ複製システムにおいて、第一のストレージから第二のストレージに転送される

データの中継する中継装置であって、第一のストレージが災害によって稼働できない状態になっても、稼働を継続できるとあらかじめ算定された位置に設置された中継装置を備え、第一のストレージは、データ転送の制御を行うデータ転送処理手段を含み、データ転送処理手段は、中継装置に対してデータ転送を完了したときに、第二のストレージに対してデータ転送を完了したとみなすことを特徴とする。

【 0 0 1 9 】

中継装置は、第一のストレージから受信したコマンドおよびデータを記憶するための不揮発性記憶手段と、データの中継制御を行う中継処理手段とを含み、中継処理手段は、第一のストレージから受信したコマンドおよびデータを不揮発性記憶手段に記憶し、記憶されたコマンドおよびデータを任意のタイミングで第二のストレージに送信する構成であってもよい。そのような構成によれば、中継装置と第二のストレージとが接続される通信ネットワークの運用コストを低減することができる。

【 0 0 2 0 】

複数の中継装置が設けられ、第一のストレージにおけるデータ転送処理手段は、第一のストレージ内のデータを同時に複数の中継装置に向けて送信してもよい。そのような構成によれば、データの転送をより迅速に行うことができる。

【 0 0 2 1 】

また、本発明による中継装置は、第一のストレージから第二のストレージに転送されるデータの中継する中継装置であって、第一のストレージから受信したデータを記憶するための記憶手段と、データの中継制御を行う中継処理手段とを含み、中継処理手段は、第一のストレージから受信したデータを記憶手段に記憶し、記憶手段にデータを記憶したときに第一のストレージに対する応答を送信し、第二のストレージに対して記憶手段に記憶したデータを送信することを特徴とする。

【 0 0 2 2 】

また、本発明によるデータを送信する送信元からデータを受信する送信先にデータを送信するデータ送受信方法において、送信元では、送信される元データか

ら少なくとも1つのエラー訂正のための冗長データを作成し、元データと冗長データとを別々のデータ送信単位で送信することを特徴とする。

【0023】

送信先では、元データと冗長データとの集合であるデータ群のすべてについて受信を完了する前に、元データについて部分的にエラー訂正処理を実行できるデータ群の一部を受信した段階で、エラー訂正処理を実行することが好ましい。そのような方法によれば、送信過程でデータの一部が廃棄されても、送信元はデータを再度送信する必要がある。

【0024】

例えば、送信元では、元データを分割データに分割し、それらの分割データのうちの1つまたは複数が消失しても元データを復元可能な冗長データを作成する。

【0025】

例えば、冗長データとしてパリティデータまたはECC (Error Correcting Code) を用いればよい。

【0026】

また、冗長データとして送信データの複製データを用いてもよい。

【0027】

元データと冗長データとを、別々の通信ネットワークに送出してもよい。そのような方法によれば、一方の通信ネットワークで障害等が発生しても、もう一方の通信ネットワークから受信したデータによって処理を進めることができる。

【0028】

また、本発明によるデータ複製システムは、第一のストレージ内のデータを第二のストレージに通信ネットワークを介してミラーリングまたはバックアップするデータ複製システムにおいて、第一のストレージは、データ転送の制御を行うデータ転送処理手段と、送信される元データから少なくとも1つのエラー訂正のための冗長データを作成する冗長化手段とを含み、データ転送処理手段は、元データと冗長化手段が作成した冗長データとを別々のデータ送信単位で送信することを特徴とする。

【 0 0 2 9 】

第二のストレージは、第一のストレージから受信した冗長データを用いてエラー訂正処理を行うデータ復元手段と、データ復元手段が復元したデータを記憶媒体に格納する格納処理手段とを含み、データ復元手段は、第一のストレージから元データと冗長データとの集合であるデータ群のすべてについて受信を完了する前に、元データについて部分的にエラー訂正処理を実行できるデータ群の一部を受信した段階で、エラー訂正処理を実行することが好ましい。そのような構成によれば、送信過程でデータの一部が廃棄されても、送信元はデータを再度送信する必要がない。

【 0 0 3 0 】

例えば、第一のストレージにおける冗長化手段は、元データを分割データに分割し、それらの分割データのうちの1つまたは複数が消失しても元データを復元可能な冗長データを作成する。

【 0 0 3 1 】

冗長化手段は、冗長データとしてパリティデータまたはECCを用いてもよい。

【 0 0 3 2 】

冗長化手段は、冗長データとして、元データの複製データを作成してもよい。

【 0 0 3 3 】

データ転送処理手段は、元データと冗長データとを、別々の通信ネットワークに送出してもよい。そのような構成によれば、一方の通信ネットワークで障害等が発生しても、もう一方の通信ネットワークから受信したデータによって処理を進めることができる。

【 0 0 3 4 】

また、本発明によるデータ複製システムは、稼働系のストレージ内のデータを通信ネットワークを介して待機系のストレージにミラーリングするデータ複製システムにおいて、稼働系のストレージは、データの書き込み要求が発生すると、書き込み対象のデータと遅延書き込み要求とを待機系のストレージに送信する遅延書き込み要求手段と、そのデータの状態であればアプリケーションがそのまま

動作を再開可能な再開可能ポイントであることを知らせるための再開可能ポイント通知を上位装置から受けると遅延書き込み実行要求を待機系のストレージに送信する書き込み実行要求手段とを含み、待機系のストレージは、データを一時的に記憶する一時記憶手段と、受信したデータを遅延書き込み要求に応じて一時記憶手段に格納するとともに、遅延書き込み実行要求を受信すると一時記憶手段に格納されているデータを記憶媒体に格納する格納処理手段とを含むことを特徴とする。

【 0 0 3 5 】

稼働系のストレージにおける遅延書き込み要求手段と遅延書き込み実行要求手段とは、遅延書き込み要求と遅延書き込み実行要求とを待機系のストレージに非同期に送信し、待機系のストレージにおける格納処理手段は、一の遅延書き込み実行要求を受信すると、一つ前の遅延書き込み実行要求と一の遅延書き込み実行要求との間に送信された遅延書き込み要求に対応するデータを全て一時記憶手段に格納し終え、一つ前の遅延書き込み実行要求以前に送信されたデータを記憶媒体に格納し終えたときに、一時記憶手段に格納されているデータを記憶媒体に格納する。

【 0 0 3 6 】

待機系のストレージにおける格納処理手段は、稼働系に異常が生ずると、一時記憶手段に格納されているデータを破棄する。

【 0 0 3 7 】

また、本発明によるデータ複製システムは、稼働系のストレージ内のデータを通信ネットワークを介して待機系のストレージにミラーリングするデータ複製システムにおいて、稼働系のストレージは、データの書き込み要求が発生すると、書き込み対象のデータと書き込み要求とを待機系のストレージに送信する書き込み要求手段と、そのデータの状態であればアプリケーションがそのまま動作を再開可能な再開可能ポイントであることを知らせるための再開可能ポイント通知を上位装置から受けるとスナップショット作成要求を待機系のストレージに送信するスナップショット作成要求手段とを含み、待機系のストレージは、書き込み要求を受信したら書き込み要求に対応するデータを書き込むべき領域を割り当てて

記憶媒体にデータを格納するとともに、記憶媒体へのデータ格納状況を示す格納情報を更新し、スナップショット作成要求を受信したらスナップショットを作成するスナップショット作成手段を含むことを特徴とする。

【 0 0 3 8 】

例えば、書き込み要求手段は、待機系ストレージのスナップショット作成手段がスナップショットを作成している場合には、スナップショット作成手段がスナップショット作成を完了した後に、書き込み要求を待機系のストレージに送信する。

【 0 0 3 9 】

また、稼働系のストレージにおける書き込み要求手段とスナップショット作成要求手段とは、書き込み要求とスナップショット作成要求とを待機系のストレージに非同期に送信し、待機系のストレージにおけるスナップショット作成手段は、各書き込み要求を受信したときに、受信した書き込み要求の直前のスナップショット作成要求に基づくスナップショットの作成を完了するまで、書き込み要求に対応するデータを記憶媒体に格納するのを待ち、一のスナップショット作成要求を受信すると、一つ前のスナップショット作成要求に基づくスナップショットの作成を完了し、一つ前のスナップショット作成要求と一のスナップショットとの間に送信された書き込み要求に対応するデータを全て記録媒体に格納し終えたときに、スナップショットを作成してもよい。

【 0 0 4 0 】

待機系のストレージにおけるスナップショット作成手段は、稼働系に異常が生ずると、直前のスナップショット作成後にデータが格納された記憶媒体の領域を未使用状態として解放し、格納情報を直前のスナップショット作成時の状態に戻す。

【 0 0 4 1 】

また、稼働系のストレージを使用する稼働系の上位装置は、再開可能ポイントになると稼働系のストレージに再開可能ポイント通知を送信する再開可能ポイント通知手段を含み、待機系のストレージを使用する待機系の上位装置は、稼働系の異常を検出すると、異常の発生を待機系のストレージに通知して待機系のスト

レージの状態を再開可能ポイントに対応する状態にするように促し、待機系のストレージの状態が再開可能ポイントに対応する状態になったときに処理を再開する構成であってもよい。

【 0 0 4 2 】

また、稼働系のストレージを使用する稼働系の上位装置は、待機系のストレージを使用する待機系の上位装置に、稼働系の処理実行状態を示す実行イメージを転送する実行イメージ転送手段と、実行イメージ転送手段が実行イメージを転送するタイミングで、稼働系のストレージに再開可能ポイント通知を送信する任意時点再開可能ポイント通知手段とを含み、待機系のストレージを使用する待機系の上位装置は、実行イメージ転送手段が転送する実行イメージを保存する実行イメージ保存手段を含み、実行イメージ転送手段は、任意の時点における実行イメージを待機系の上位装置に転送し、待機系のストレージを使用する待機系の上位装置は、稼働系の異常を検出すると、異常の発生を待機系のストレージに通知して、待機系のストレージの状態を実行イメージの転送タイミングに対応する状態にするように促し、待機系のストレージの状態が実行イメージの転送タイミングに対応する状態になった場合に実行イメージを用いて処理を再開する構成であってもよい。

【 0 0 4 3 】

稼働系の上位装置における実行イメージ転送手段は、前回転送した実行イメージから変更された部分のみを転送してもよい。

【 0 0 4 4 】

また、本発明によるストレージ内のデータを複製するためのプログラムは、第一のストレージ内のデータを通信ネットワークおよび中継装置を介して第二のストレージにミラーリングまたはバックアップするデータ複製システムにおける中継装置内に設けられているコンピュータに、第一のストレージから受信したデータを中継装置内の記憶媒体に記憶する処理と、データを中継装置内の記憶媒体に記憶したときに第一のストレージに対する応答を送信する処理と、中継装置内の記憶媒体に記憶したデータを第二のストレージに送信する処理とを実行させる。

【 0 0 4 5 】

また、本発明によるストレージ内のデータを複製するためのプログラムは、第一のストレージ内のデータを第二のストレージに通信ネットワークを介してミラーリングまたはバックアップするデータ複製システムにおける第一のストレージ内に設けられているコンピュータに、送信される元データから少なくとも1つのエラー訂正のための冗長データを作成する処理と、元データと冗長データとを別々のデータ送信単位で送信する処理とを実行させる。

【 0 0 4 6 】

また、本発明によるストレージ内のデータを複製するためのプログラムは、稼働系のストレージ内のデータを通信ネットワークを介して待機系のストレージにミラーリングするデータ複製システムにおける稼働系のストレージ内に設けられているコンピュータに、

データの書き込み要求が発生すると、書き込み対象のデータと一時記憶装置にデータを格納することを指示する遅延書き込み要求とを待機系のストレージに送信する処理と、そのデータの状態であればアプリケーションがそのまま動作を再開可能なポイントであることを知らせるための再開可能ポイント通知を上位装置から受けると、一時記憶装置に格納されているデータを記憶媒体に格納することを指示する遅延書き込み実行要求を待機系のストレージに送信する処理とを実行させる。

【 0 0 4 7 】

また、本発明によるストレージ内のデータを複製するためのプログラムは、稼働系のストレージ内のデータを通信ネットワークを介して待機系のストレージにミラーリングするデータ複製システムにおける稼働系のストレージ内に設けられているコンピュータに、

データの書き込み要求が発生すると、書き込み対象のデータと記憶媒体にデータを書き込むことを指示する書き込み要求とを待機系のストレージに送信する処理と、そのデータの状態であればアプリケーションがそのまま動作を再開可能なポイントであることを知らせるための再開可能ポイント通知を上位装置から受けるとスナップショット作成要求を待機系のストレージに送信する処理を実行させる。

【 0 0 4 8 】

【発明の実施の形態】

以下、本発明の実施の形態を図面を参照して説明する。

【 0 0 4 9 】

実施の形態 1.

図 1 は、本発明によるデータ複製システムの第 1 の実施の形態を示すブロック図である。図 1 に示すデータ複製システムにおいて、ストレージ 1 1 が、ストレージ 1 1 を使用するホスト（ホストコンピュータ）1 0 とローカルに接続されている。ストレージ 1 1 は、インターネットや専用線等の通信ネットワーク（以下、ネットワークという。）1 3 を介して中継装置 1 5 に接続されている。また、中継装置 1 5 は、ネットワーク 1 4 を介してストレージ 1 2 に接続されている。

【 0 0 5 0 】

ストレージ 1 1, 1 2 は、例えば、単体の磁気ディスク装置、光ディスク装置または光磁気ディスク装置である。ストレージ 1 1, 1 2 として、単体の磁気ディスク装置、光ディスク装置または光磁気ディスク装置の集合であるディスクアレイ装置を使用することもできる。ホスト 1 0 とストレージ 1 1 とは、SCSI、ファイバチャネル（Fibre channel）、イーサネット（登録商標）等で接続される。

【 0 0 5 1 】

図 2 は、図 1 に示すストレージ 1 1 の構成例を示すブロック図である。なお、ストレージ 1 2 の構成も、図 2 に示すような構成である。図 2 に示すように、ストレージ 1 1 は、ストレージコントローラ 1 0 0 とストレージ本体である記憶媒体 1 0 1 とを含む。ストレージコントローラ 1 0 0 は、ホスト 1 0 および他のストレージまたは中継装置と通信を行う通信部 1 0 2、各処理のシーケンスを管理する処理シーケンサ 1 0 3、記憶媒体 1 0 1 に対する処理命令の順序制御を行う I/O スケジューラ 1 0 4、I/O スケジューラ 1 0 4 が発行する処理命令に従って記憶媒体 1 0 1 の動作を制御する媒体処理部 1 0 5、およびホスト 1 0 等から記憶媒体 1 0 1 へのデータおよび記憶媒体 1 0 1 からホスト 1 0 等へのデータを一時記憶するバッファメモリ 1 0 6 を含む。処理シーケンサ 1 0 3 は、例えば、プ

プログラムに従って動作するCPUで実現される。

【0052】

図2では、媒体制御部105と記憶媒体101との組み合わせを一組だけ含む場合を示した。ストレージ11、12が複数の記録媒体101を含み、それぞれの記憶媒体に対応して媒体制御部105が設けられていてもよい。ただし、個々のストレージ11、12に含まれるI/Oスケジューラ104は一つである。記憶媒体101が複数存在する場合、処理シーケンサ103は処理命令の対象となる記憶媒体を指定し、I/Oスケジューラ104は、指定された記憶媒体に対応する媒体制御部105に処理を行わせる。

【0053】

図3は、図1に示す中継装置15の構成例を示すブロック図である。中継装置15は、ストレージ11、12と通信を行う通信部150、中継処理のシーケンス管理を行う中継処理部151、およびストレージ11、12から受信したデータを一時記憶するバッファメモリ152を含む。中継処理部151は、例えば、プログラムに従って動作するCPUで実現される。

【0054】

図1および図2に示すように、ストレージ11には通信部102が設けられ、ストレージ11自身が主体的にデータを転送する。すなわち、ホスト10がストレージ11からデータを読み出して中継装置15にデータを送信するのではなく、ストレージ11が直接中継装置15にデータを送信する。ストレージ12も、中継装置15から直接データを受信する。ただし、ストレージ11が中継装置15にデータを送信するタイミングは、ホスト10が指示する。

【0055】

また、中継装置15は、ストレージ11、12の設置場所において地震等の災害が発生したときに、災害の影響が波及すると想定される範囲の外に設置される。すなわち、ストレージ11、12が災害によって稼働できない状態になっても、その位置であれば稼働を継続できる位置（ストレージ11、12からの距離）があらかじめ算定され、中継装置15は、算定された位置に設置されている。例えば、ストレージ11があらかじめ想定した震度（例えば震度6～7）の地震に

被災したときに、中継装置 1 5 には破損等が生じない程度の震度になるような位置に、中継装置 1 5 が設置される。さらに、ストレージ 1 1 またはストレージ 1 2 と中継装置 1 5 との間のデータ転送時間が、ストレージ 1 1 とストレージ 1 2 との間で直接データを転送した場合のデータ転送時間よりも短くなるように、中継装置 1 5 が設置される。よって、データ複製システムは、耐災害データ管理システムとして機能する。

【 0 0 5 6 】

次に、データ複製システムの動作を、図 4 および図 5 のフローチャートを参照して説明する。ここでは、バックアップ処理として、データの転送元としてのストレージ 1 1 が、データの転送先としてのストレージ 1 2 に向けてデータを転送する場合を例にする。図 4 は、ストレージコントローラ 1 0 0 における処理シーケンサ 1 0 3 の動作を示すフローチャートであり、図 5 は中継装置 1 5 における中継処理部 1 5 1 の動作を示すフローチャートである。

【 0 0 5 7 】

ホスト 1 0 は、バックアップ処理を実行するときに、ストレージ 1 1 のストレージコントローラ 1 0 0 における通信部 1 0 2 に対して複写コマンドを出力する。複写コマンドは、バックアップを指示するコマンドである。複写コマンドには、複写対象のデータの範囲すなわちバックアップ処理対象のデータ範囲を指定する情報および複写先のストレージを指定する情報が含まれている。通信部 1 0 2 は、複写コマンドを受け取ると、複写コマンドを処理シーケンサ 1 0 3 に渡し、複写処理を開始することを処理シーケンサ 1 0 3 に指示する。

【 0 0 5 8 】

処理シーケンサ 1 0 3 は、複写コマンドを受け取ると、図 4 に示すように、まず、複写対象のデータの範囲の先頭のブロックをデータ転送対象のブロックとして設定する（ステップ S 1 0 0）。次いで、通信部 1 0 2 に、複写コマンドで指定された複写先のストレージに対応した中継装置に対して write コマンドを送信するように指示する（ステップ S 1 0 1）。write コマンドには、ブロックサイズすなわちデータ量を示す情報が含まれている。また、データ転送対象のデータの読み出し要求を I O スケジューラ 1 0 4 に登録する。I O スケジューラ 1 0 4

は、読み出し要求に応じて、媒体制御部 1 0 5 に、データ転送対象のデータの読み出し指示を行う。媒体制御部 1 0 5 は、読み出し指示に従って、データ転送対象のデータを記憶媒体 1 0 1 からバッファメモリ 1 0 6 に出力させる（ステップ S 1 0 2）。媒体制御部 1 0 5 は、データ転送対象の全てのデータが記憶媒体 1 0 1 からバッファメモリ 1 0 6 に出力されると、読み出し完了通知を処理シーケンサ 1 0 3 に出力する。

【 0 0 5 9 】

そして、処理シーケンサ 1 0 3 は、通信部 1 0 2 を介して入力される中継装置 1 5 からの受信準備完了のメッセージと、媒体制御部 1 0 5 からの読み出し完了通知との双方を待ち（ステップ S 1 0 3）。双方が入力されたら、バッファメモリ 1 0 6 に記憶されたデータを中継装置 1 5 に転送するように通信部 1 0 2 に指示する。通信部 1 0 2 は、指示に応じてバッファメモリ 1 0 6 に記憶されたデータを中継装置 1 5 に送信する（ステップ S 1 0 4）。そして、処理シーケンサ 1 0 3 は、通信部 1 0 2 を介して入力される中継装置 1 5 からの受信完了のメッセージを待つ（ステップ S 1 0 5）。

【 0 0 6 0 】

通信部 1 0 2 は、バッファメモリ 1 0 6 に記憶された全てのデータを中継装置 1 5 に送信し、中継装置 1 5 から受信完了のメッセージを受けたら、受信完了のメッセージを処理シーケンサ 1 0 3 を出力する。処理シーケンサ 1 0 3 は、受信完了のメッセージを入力すると、複写対象の全てのデータの中継装置 1 5 への転送が完了したか否か確認する（ステップ S 1 0 6）。完了していなければ、複写対象のデータの範囲の次のブロックをデータ転送対象のブロックとして設定し（ステップ S 1 0 8）、ステップ S 1 0 1 に戻る。

【 0 0 6 1 】

複写対象の全てのデータの転送が完了していれば、処理シーケンサ 1 0 3 は、ホスト 1 0 に対して完了通知を出力するように通信部 1 0 2 に指示し（ステップ S 1 0 7）、処理を終了する。

【 0 0 6 2 】

なお、1 ブロックのデータ量は、あらかじめシステムに設定されている量であ

る。また、処理シーケンサ 1 0 3 は、1 ブロックのデータ量を、転送先のストレージ 1 2 のバッファメモリ 1 0 6 および中継装置 1 5 のバッファメモリ 1 5 2 の容量に応じて変化させるようにしてもよい。

【 0 0 6 3 】

処理シーケンサ 1 0 3 が I O スケジューラ 1 0 4 に登録するものとして、処理の種類（読み出し／書き込み）と、処理を識別するための I D と、処理の対象となる記憶媒体 1 0 1 中の領域を示す情報と、処理の対象となるバッファメモリ 1 0 6 の領域を示す情報とがある。なお、処理シーケンサ 1 0 3 が複数の読み出し処理命令や複数の書き込み処理命令を指示する場合もあり、処理を識別するための I D は、このような場合に各処理を識別するために用いられる。従って、図 4 に示された処理では、処理シーケンサ 1 0 3 は、データ転送対象のデータの読み出し要求を登録する際に、処理の種類、読み出し処理を識別するための I D を登録するとともに、処理の対象となる記憶媒体 1 0 1 中の領域としてデータ転送対象のブロックを登録する。また、記憶媒体が複数ある場合には、処理シーケンサ 1 0 3 は、記憶媒体を特定するための情報も登録する。そして、処理シーケンサ 1 0 3 は、登録時に指定した I D によって、どの処理が完了したのかを判別する。

【 0 0 6 4 】

I O スケジューラ 1 0 4 は、処理シーケンサ 1 0 3 によって登録された処理を記録する。そして、記録されている処理をあらかじめ決められているアルゴリズムに従って選択し、取り出した処理を媒体制御部 1 0 5 に実行させる。あらかじめ決められているアルゴリズムとして、例えば、登録された順がある。また、記憶媒体 1 0 1 の磁気ヘッドや光ヘッドの現在位置から処理対象の位置までの移動距離が最も小さくなる処理を最初を選択してもよい。また、ストレージ 1 1 が複数の記録媒体 1 0 1 を含むディスクアレイ装置であってそれぞれの記憶媒体に対応して媒体制御部 1 0 5 が設けられている場合には、I O スケジューラ 1 0 4 は、処理を取り出そうとした記憶媒体 1 0 1 に対応した媒体制御部 1 0 5 が担当する記憶媒体 1 0 1 を対象とした処理を選択する。

【 0 0 6 5 】

媒体制御部 1 0 5 の実行中の処理が完了すると、I O スケジューラ 1 0 4 は次の処理を取り出す。処理が読み出し処理であった場合には、媒体制御部 1 0 5 は、記憶媒体 1 0 1 に、記憶媒体 1 0 1 における指定された領域からデータを読み出させて、読み出したデータをバッファメモリ 1 0 6 の指定された領域に書き込ませる。処理が書き込み処理であった場合には、媒体制御部 1 0 5 は、記憶媒体 1 0 1 に、バッファメモリ 1 0 6 の指定された領域のデータを、記憶媒体 1 0 1 における指定された領域に書き込ませる。そして、媒体制御部 1 0 5 は、処理が完了したときには、処理シーケンサ 1 0 3 に処理の完了を通知する。

【 0 0 6 6 】

通信部 1 0 2 は、外部から入力されたデータが、コマンド、受信準備完了のメッセージ、受信完了のメッセージ等の制御系メッセージであった場合には、入力されたデータを処理シーケンサ 1 0 3 に渡す。また、ストレージ 1 1 に書き込まれるべきデータであった場合には、データを格納すべき場所を処理シーケンサ 1 0 3 に問い合わせる。そして、処理シーケンサ 1 0 3 から指定されたバッファメモリ 1 0 6 中の領域にデータを格納する。

【 0 0 6 7 】

また、通信部 1 0 2 は、処理シーケンサ 1 0 3 から指定されたコマンドまたは完了通知を、指定された中継装置またはホストに向けて送信する処理も行う。さらに、処理シーケンサ 1 0 3 から指定されたバッファメモリ 1 0 6 中のデータを、指定された中継装置またはホストに向けて送信する処理も行う。

【 0 0 6 8 】

なお、ホスト 1 0 は、ストレージ 1 1 から完了通知を受けたら、すなわち、中継装置 1 5 へのデータ転送が完了したら、ストレージ 1 1 からストレージ 1 2 へのデータ転送が完了したと認識する。

【 0 0 6 9 】

次に、中継装置 1 5 の動作を説明する。中継装置 1 5 において、ストレージ 1 1, 1 2 からのデータ（コマンドまたは転送されるデータ）は、通信部 1 5 0 で受信される。通信部 1 5 0 は、ストレージ 1 1 がステップ S 1 0 1 において送信した write コマンドを受信すると、write コマンドを中継処理部 1 5 1 に渡す。

【 0 0 7 0 】

中継処理部 1 5 1 は、図 5 に示すように、ストレージ 1 1 から送られてくるデータを格納するのに必要な領域をバッファメモリ 1 5 2 に確保する（ステップ S 1 2 0）。中継処理部 1 5 1 は、write コマンドに含まれるデータ量の情報に基づいてデータの格納領域を確保する。そして、受信準備完了の通知をストレージ 1 1 に送るように通信部 1 5 0 に指示する（ステップ S 1 2 1）。通信部 1 5 0 は、指示に応じて、ストレージ 1 1 に受信準備完了のメッセージを送信する。また、write コマンドをストレージ 1 2 に送るように通信部 1 5 0 に指示する（ステップ S 1 2 2）。通信部 1 5 0 は、指示に応じて、ストレージ 1 2 に write コマンドを送信する。この write コマンドは、ストレージ 1 1 から受信するデータをストレージ 1 2 に書き込ませるための write コマンドである。

【 0 0 7 1 】

次いで、ストレージ 1 1 からデータが届くのを待ち（ステップ S 1 2 3）、データが届いて通信部 1 5 0 からデータを格納すべきバッファメモリ 1 5 2 の領域の問い合わせを受けると、ステップ S 1 2 0 で確保した領域を通信部 1 5 0 に知らせる（ステップ S 1 2 4）。また、データのバッファメモリ 1 5 2 への格納の完了を待ち（ステップ S 1 2 5）、全てのデータがバッファメモリ 1 5 2 に格納されたことが通信部 1 5 0 から通知されると、ストレージ 1 1 に write コマンドに対する完了を通知するように通信部 1 5 0 に指示する（ステップ S 1 2 6）。通信部 1 5 0 は、指示に応じて、ストレージ 1 1 に受信完了のメッセージを送信する。

【 0 0 7 2 】

そして、ストレージ 1 2 から準備完了のメッセージ（ステップ S 1 2 2 において送信した write コマンドに対する応答）が送信されるのを待つ（ステップ S 1 2 7）。ストレージ 1 2 からの準備完了のメッセージを受信したことが通信部 1 5 0 から通知されると、通信部 1 5 0 に、バッファメモリ 1 5 2 に格納されたデータをストレージ 1 2 に向けて送信させる（ステップ S 1 2 8）。その後、ストレージ 1 2 から受信完了のメッセージが送信されるのを待ち（ステップ S 1 2 9）、ストレージ 1 2 からの受信完了のメッセージを受信したことが通信部 1 5 0

から通知されると、処理を終了する。

【0073】

通信部150は、外部からコマンドを受信すると、受信したコマンドを中継処理部151に渡す。また、外部から各種のメッセージを受信すると、受信したことを中継処理部151に通知する。さらに、データを受信した場合には、データを格納すべきバッファメモリ152の領域を中継処理部151に問い合わせ、バッファメモリ152中の指定された領域にデータを格納する。また、中継処理部151の指示に応じて、指定されたメッセージを指定されたストレージに送信する。さらに、バッファメモリ152中の領域と送信先のストレージとが指定された場合には、その領域中のデータを指定されたストレージに送信する。

【0074】

中継処理部151がデータの送信先のストレージを決定する際に、あらかじめ送信先のストレージ（この例ではストレージ12）が固定的に決められている場合には特に選択処理を行わないが、データの転送元のストレージ（この例ではストレージ11）に応じてデータの送信先のストレージが決まる場合には、転送元のストレージに応じて送信先のストレージを選択する処理を行う。また、送信先のストレージのデータを送信するのに先立って、データの転送元のストレージから送信先が指定されることもある。なお、バッファメモリ152に格納されたデータは、送信先のストレージ（この例ではストレージ12）へのデータの送信が完了すると破棄される。

【0075】

ストレージ12の通信部102は、中継装置15がステップS122において送信したwrite コマンドを受信すると、write コマンドを処理シーケンサ103に渡す。処理シーケンサ103は、中継装置15から送られてくるデータを格納するのに必要な領域をバッファメモリ106に確保する。そして、通信部102に、準備完了のメッセージを中継装置15に対して送信させる。この準備完了メッセージは、中継装置15がステップS127において待つメッセージである。その後、中継装置15からデータが送られてくると、ストレージ12の通信部102は、データを格納すべき領域を処理シーケンサ103に問い合わせ、バッ

ァメモリ106に格納する。データを全て格納したならば、受信完了のメッセージを中継装置15に送信する。この受信完了メッセージは、中継装置15がステップS129において待つメッセージである。また、ストレージ12の処理シーケンサ103は、バッファメモリ106へのデータ格納後、I/Oスケジューラ104に対して、処理の種類（この場合には書き込み）と、処理の識別IDと、処理の対象となる記憶媒体101中の領域を示す情報と、処理の対象となるバッファメモリ106の領域を示す情報とを登録する。I/Oスケジューラ104は、書き込み要求に応じて、媒体制御部105に、書き込み対象のデータの書き込み指示を行う。媒体制御部105は、登録内容に応じてバッファメモリ106から記憶媒体101へのデータの書き込み処理を行う。

【0076】

この実施の形態では、中継装置15は、ストレージ11の設置場所において地震等の災害が発生したときに、災害の影響が波及すると想定される範囲の外に設置される。また、中継装置15は、ストレージ11と中継装置15との間のデータ転送時間が、ストレージ11とストレージ12との間で直接データを転送した場合のデータ転送時間よりも短くなるような位置に設置されている。さらに、バックアップのためにストレージ11からストレージ12に向けてデータを転送する際に、ストレージ11を使用しているホスト10は、中継装置15へのデータ転送が完了したら、ストレージ11からストレージ12へのデータ転送が完了したと認識する。

【0077】

従って、ストレージ11が被災しても中継装置15は被災せず、かつ、ストレージ11から中継装置15へのデータ転送時間が短いので、耐障害性が向上する。また、ホスト10は、ストレージ12にデータが格納されるのを待たずに、次の処理を開始することができるので、データ転送に伴う処理の遅れが改善される。

【0078】

実施の形態2.

第1の実施の形態では、中継装置15は、ストレージ11からの1ブロックの

データを全て受信してから、ストレージ 1 2 に対するデータの送信を開始したが、ストレージ 1 1 からの 1 ブロックのデータの受信の完了を待たずに、ストレージ 1 2 に対するデータの送信を開始してもよい。図 6 は、ストレージ 1 1 からのデータの受信の完了を待たずにストレージ 1 2 に対するデータの送信を開始する制御を行う第 2 の実施の形態の中継処理部 1 5 1 の動作を示すフローチャートである。なお、データ複製システムの構成およびストレージ 1 1, 1 2 と中継装置 1 5 の構成は第 1 の実施の形態の場合と同じである（図 1 ～図 3 参照）。

【 0 0 7 9 】

中継処理部 1 5 1 は、第一の実施の形態に示す場合と同様に、ストレージ 1 1 がステップ S 1 0 1 において送信した write コマンドを受信する。すると、中継処理部 1 5 1 は、図 6 に示すように、ストレージ 1 1 から送られてくるデータを格納するのに必要な領域をバッファメモリ 1 5 2 に確保する（ステップ S 1 4 0）。中継処理部 1 5 1 は、ストレージ 1 1 からの write コマンドに含まれるデータ量の情報に基づいてデータの格納領域を確保する。そして、受信準備完了の通知をストレージ 1 1 に送るように通信部 1 5 0 に指示する（ステップ S 1 4 1）。また、write コマンドをストレージ 1 2 に送るように通信部 1 5 0 に指示する（ステップ S 1 4 2）。この write コマンドは、ストレージ 1 1 から受信するデータをストレージ 1 2 に書き込ませるための write コマンドである。

【 0 0 8 0 】

次いで、ストレージ 1 1 からデータが届くのを待ち（ステップ S 1 4 3）、データが届いて通信部 1 5 0 からデータを格納すべきバッファメモリ 1 5 2 の領域の問い合わせを受けると、ステップ S 1 4 0 で確保した領域を通信部 1 5 0 に知らせる（ステップ S 1 4 4）。次いで、ストレージ 1 2 から準備完了のメッセージが送信されるのを待ち（ステップ S 1 4 5）、ストレージ 1 2 からの準備完了のメッセージを受信したことが通信部 1 5 0 から通知されると、通信部 1 5 0 に、バッファメモリ 1 5 2 に格納されたデータをストレージ 1 2 に向けて送信させる（ステップ S 1 4 6）。

【 0 0 8 1 】

ステップ S 1 4 6 の処理を行っているときに、ストレージ 1 1 からのデータが

バッファメモリ 1 5 2 に格納され、バッファメモリ 1 5 2 から読み出されたデータがストレージに向けて送信されるが、通信部 1 5 0 は、データの読み出し位置（読み出しアドレス）が格納位置（格納アドレス）を追い越さないように制御する。すなわち、読み出し位置が格納位置に追いついたら、バッファメモリ 1 5 2 からのデータの読み出しを中止する。

【 0 0 8 2 】

そして、データのバッファメモリ 1 5 2 への格納が完了するのを待ち（ステップ S 1 4 7）、全てのデータがバッファメモリ 1 5 2 に格納されたことが通信部 1 5 0 から通知されると、ストレージ 1 1 に write コマンドに対する完了を通知するように通信部 1 5 0 に指示する（ステップ S 1 4 8）。その後、ストレージ 1 2 から受信完了のメッセージが送信されるのを待ち（ステップ S 1 4 9）、ストレージ 1 2 からの受信完了のメッセージを受信したことが通信部 1 5 0 から通知されると、処理を終了する。

【 0 0 8 3 】

この実施の形態では、第 1 の実施の形態に比べて、中継装置 1 5 からストレージ 1 2 へのデータの送信を早めに完了させることができる。なお、ストレージ 1 1、1 2 の動作は、第 1 の実施の形態のそれらの動作と同じである。

【 0 0 8 4 】

実施の形態 3.

第 1 および第 2 の実施の形態では、ストレージ 1 1 のデータのバックアップが実現されたが、ミラーリングによってストレージ 1 1 のデータをストレージ 1 2 に転送するようにしてもよい。図 7 は、第 3 の実施の形態、すなわちミラーリングを行う場合のストレージ 1 1 のストレージコントローラ 1 0 0 における処理シーケンス 1 0 3 の動作を示すフローチャートである。なお、データ複製システムの構成およびストレージ 1 1、1 2 と中継装置 1 5 の構成は第 1 の実施の形態の場合と同じである（図 1 ～図 3 参照）。また、ストレージ 1 1 において、通信部 1 0 2、I/O スケジューラ 1 0 4 および媒体制御部 1 0 5 の動作は、第 1 の実施の形態のそれらの動作と同じである。

【 0 0 8 5 】

第 1 および第 2 の実施の形態に示したバックアップは、ホスト 1 0 がストレージ 1 1 に対して複写コマンドを出力したことを契機に開始される。第 3 の実施の形態として示すミラーリングは、ホスト 1 0 がストレージ 1 1 に対してデータの書き込みを指示する write コマンドを出力したことを契機に開始される。

【 0 0 8 6 】

ストレージ 1 1 の通信部 1 0 2 がホスト 1 0 から write コマンドを受信すると、通信部 1 0 2 はその write コマンドを処理シーケンサ 1 0 3 に渡す。すると、処理シーケンサ 1 0 3 は、図 7 に示すように、ホスト 1 0 から受け取るデータを格納するのに必要な領域をバッファメモリ 1 0 6 に確保する（ステップ S 1 6 0）。また、準備完了の通知をホスト 1 0 に送るように通信部 1 0 2 に指示する（ステップ S 1 6 1）。通信部 1 0 2 は、指示に応じて、準備完了の通知をホスト 1 0 に送る。

【 0 0 8 7 】

そして、処理シーケンサ 1 0 3 は、ホスト 1 0 からデータが到着するのを待ち（ステップ S 1 6 2）、データが届いて通信部 1 0 2 からデータを格納すべきバッファメモリ 1 0 6 の領域の問い合わせを受けると、ステップ S 1 6 0 で確保した領域を通信部 1 0 2 に知らせる（ステップ S 1 6 3）。次いで、write コマンドを中継装置 1 5 に送るように通信部 1 0 2 に指示する（ステップ S 1 6 4）。通信部 1 0 2 は、指示に応じて、中継装置 1 5 に write コマンドを送信する。

【 0 0 8 8 】

次いで、処理シーケンサ 1 0 3 は、ホスト 1 0 からのデータのバッファメモリ 1 0 6 への格納の完了を待ち（ステップ S 1 6 5）、全てのデータがバッファメモリ 1 0 6 に格納されたことが通信部 1 0 2 から通知されると、処理シーケンサ 1 0 3 は、I/Oスケジューラ 1 0 4 に対して、処理の種類（この場合には書き込み）と、処理の識別 ID と、処理の対象となる記憶媒体 1 0 1 中の領域を示す情報と、処理の対象となるバッファメモリ 1 0 6 の領域を示す情報とを登録する。

I/Oスケジューラ 1 0 4 は、書き込み要求に応じて、媒体制御部 1 0 5 に、書き込み対象のデータの書き込み指示を行う。媒体制御部 1 0 5 は、登録内容に応じてバッファメモリ 1 0 6 から記憶媒体 1 0 1 へのデータの書き込み処理を行う（

ステップ S 1 6 6)。

【 0 0 8 9 】

そして、処理シーケンサ 1 0 3 は、中継装置 1 5 から受信準備完了のメッセージ（ステップ S 1 6 4 において送った write コマンドに対する応答）が送信されるのを待つ（ステップ S 1 6 7）。中継装置 1 5 からの受信準備完了のメッセージを受信したことが通信部 1 0 2 から通知されると、処理シーケンサ 1 0 3 は、通信部 1 0 2 に、バッファメモリ 1 0 6 に格納されたデータを中継装置 1 5 に向けて送信させる（ステップ S 1 6 8）。その後、中継装置 1 5 から受信完了のメッセージが送信されるのと、媒体制御部 1 0 5 からの書き込み完了通知とを待ち（ステップ S 1 6 9）、中継装置 1 5 からの受信完了のメッセージを受信したことが通信部 1 0 2 から通知され、かつ、媒体制御部 1 0 5 からの書き込み完了通知を受けると、ホスト 1 0 に書き込みの完了を通知し（ステップ S 1 7 0）、処理を終了する。ホスト 1 0 は、ストレージ 1 1 から完了通知を受けたら、すなわち、中継装置 1 5 へのデータ転送および記憶媒体 1 0 1 へのデータ書き込みが完了したら、ミラーリングが完了したと認識する。

【 0 0 9 0 】

中継装置 1 5 がストレージ 1 1 から送られるデータを中継する動作や、ストレージ 1 2 が中継装置 1 5 から送られるデータを記憶媒体 1 0 1 に記憶させる動作は、第一の実施の形態と同様である。

【 0 0 9 1 】

この実施の形態でも、ストレージ 1 1 が被災しても中継装置 1 5 は被災せず、かつ、ストレージ 1 1 から中継装置 1 5 へのデータ転送時間が短いので、耐障害性が向上する。また、ホスト 1 0 が write コマンドを送信してから次の処理を開始するまでの時間を短縮化できる。

【 0 0 9 2 】

実施の形態 4.

図 8 は、本発明によるデータ複製システムの第 4 の実施の形態を示すブロック図である。図 8 に示すデータ複製システムにおいて、ストレージ 1 1 が、ストレージ 1 1 を使用するホスト 1 0 とローカルに接続されている。ストレージ 1 1 は

、ネットワーク 13 を介して中継装置 15-1 ~ 15-n に接続されている。また、中継装置 15-1 ~ 15-n は、ネットワーク 14 を介してストレージ 12 に接続されている。なお、ストレージ 11, 12 の構成は、第 1 の実施の形態のストレージ 11, 12 の構成と同じであり、中継装置 15-1 ~ 15-n の構成は、第 1 の実施の形態の中継装置 15 の構成と同じである。また、各中継装置 15-1 ~ 15-n がストレージ 11 から受信したデータをストレージ 12 に転送する動作や、ストレージ 12 が受信したデータを記憶媒体 101 に記憶させる動作は、第 1 の実施の形態の中継装置 15 やストレージ 12 の動作と同様である。

【0093】

第一の実施の形態と同様に、ストレージ 11 は主体的にデータを送信する。すなわち、ホスト 10 がストレージ 11 からデータを読み出して中継装置 15-1 ~ 15-n にデータを送信するのではなく、ストレージ 11 が直接中継装置 15 にデータを送信する。ストレージ 12 も、中継装置 15 から直接データを受信する。また、中継装置 15-1 ~ 15-n は、ストレージ 11, 12 の設置場所において地震等の災害が発生したときに、災害の影響が波及すると想定される範囲の外に設置される。さらに、ストレージ 11 またはストレージ 12 と中継装置 15-1 ~ 15-n との間のデータ転送時間が、ストレージ 11 とストレージ 12 との間で直接データを転送した場合のデータ転送時間よりも短くなるように、中継装置 15-1 ~ 15-n が設置される。

【0094】

次に、データ複製システムの動作を説明する。図 9 は、ストレージ 11 のストレージコントローラ 100 における処理シーケンサ 103 の動作を示すフローチャートである。ここでは、データの転送元としてのストレージ 11 が、データの転送先としてのストレージ 12 に向けてデータを転送して、バックアップを行う場合の例を説明する。

【0095】

処理シーケンサ 103 は、通信部 102 を介してホスト 10 から複写コマンドを受け取ると、図 9 に示すように、まず、複写対象のデータの範囲の先頭のブロックをデータ転送対象のブロックとして設定する（ステップ S200）。次いで

、中継装置 15-1 ~ 15-n のうちから使用する中継装置を選択する（ステップ S 2 0 1）。ここでは、中継装置 15-1 が選択されたとする。そして、通信部 1 0 2 に、選択した中継装置 15-1 に対して write コマンドを送信するように指示する（ステップ S 2 0 2）。また、データ転送対象のデータの読み出し要求を I O スケジューラ 1 0 4 に登録する（ステップ S 2 0 3）。ステップ S 2 0 3 において読み出し要求が登録されたときの I O スケジューラ 1 0 4 と媒体制御部 1 0 5 の動作は、第 1 の実施の形態で説明したステップ S 1 0 2 における動作と同様である。

【 0 0 9 6 】

そして、処理シーケンサ 1 0 3 は、通信部 1 0 2 を介して入力される中継装置 15-1 からの受信準備完了のメッセージと、媒体制御部 1 0 5 からの読み出し完了通知との双方を待ち（ステップ S 2 0 4）。双方が入力されたら、バッファメモリ 1 0 6 に記憶されたデータを中継装置 15-1 に転送するように通信部 1 0 2 に指示する（ステップ S 2 0 5）。そして、処理シーケンサ 1 0 3 は、通信部 1 0 2 を介して入力される中継装置 15-1 からの受信完了のメッセージを待つ（ステップ S 2 0 6）。

【 0 0 9 7 】

通信部 1 0 2 は、ステップ S 2 0 3 においてバッファメモリ 1 0 6 に記憶させたデータを中継装置 15-1 に送信し、中継装置 15-1 から受信完了のメッセージを受けたら、受信完了のメッセージを処理シーケンサ 1 0 3 に出力する。処理シーケンサ 1 0 3 は、受信完了のメッセージを入力すると、複写対象の全てのデータの中継装置 15 への転送が完了したか否か確認する（ステップ S 2 0 7）。完了していなければ、複写対象のデータの範囲の次のブロックをデータ転送対象のブロックとして設定し（ステップ S 2 0 9）、ステップ S 2 0 1 に戻る。

【 0 0 9 8 】

複写対象の全てのデータの転送が完了していれば、処理シーケンサ 1 0 3 は、ホスト 1 0 に対して完了通知を出力するように通信部 1 0 2 に指示し（ステップ S 2 0 8）、処理を終了する。

【 0 0 9 9 】

なお、1ブロックのデータ量は、あらかじめシステムに設定されている量である。また、処理シーケンサ103は、1ブロックのデータ量を、転送先のストレージ12のバッファメモリ106および中継装置15のバッファメモリ152の容量に応じて変化させるようにしてもよい。

【0100】

ステップS201での中継装置を選択する方法として、例えば、中継装置15-1～15-nを順に選択したり、負荷の軽い（他のデータ転送に用いられていない）中継装置を選択したり、乱数を用いて選択したりする方法がある。

【0101】

あるいは、複数の中継装置を同時に使用してデータ転送を並列に実行してもよい。図10は、複数の中継装置を同時に使用する場合のストレージ11の動作を示すフローチャートである。ここでは、1ブロックのデータを転送する場合を例にする。

【0102】

ストレージ11において、処理シーケンサ103が、複写対象のデータの範囲の先頭のブロックをデータ転送対象のブロックとして設定すると（ステップS220）、ストレージ11は、全中継装置15-1～15-nに対してデータを転送する（ステップS221）。

【0103】

図11は、ステップS221の処理を具体的に示すフローチャートである。処理シーケンサ103は、個々の中継装置15-1～15-nに対して、それぞれ、図11に示すフローチャートに従ってデータを送信する。ここでは、中継装置15-1にデータを送信する場合を例に説明する。処理シーケンサ103は、バッファメモリ106に格納されているデータから、中継装置15-1に送信するデータを選択する（ステップS240）。そして、その中継装置15-1に対してwrite コマンドを送るように通信部102に指示する（ステップS241）。また、IOスケジューラ104に、転送対象のデータの読み出し要求を登録する（ステップS242）。ステップS242において読み出し要求が登録されたときのIOスケジューラ104と媒体制御部105の動作は、第1の実施の形態で

説明したステップ S 1 0 2 における動作と同様である。処理シーケンサ 1 0 3 は、通信部 1 0 2 を介して入力される中継装置 1 5 - 1 からの受信準備完了のメッセージと、媒体制御部 1 0 5 からの読み出し完了通知との双方を待ち（ステップ S 2 4 3）。双方が入力されたら、バッファメモリ 1 0 6 に記憶されたデータを中継装置 1 5 - 1 に送信するように通信部 1 0 2 に指示する（ステップ S 2 4 4）。ここでは中継装置 1 5 - 1 にデータを送信する場合を例に説明したが、処理シーケンサ 1 0 3 は、ステップ S 2 4 0 ~ S 2 4 4 の処理を全ての中継装置 1 5 - 1 ~ 1 5 - n について実行する。

【 0 1 0 4 】

さらに、処理シーケンサ 1 0 3 は、いずれかの中継装置から受信完了のメッセージを受けたことを通信部 1 0 2 から通知されたら（ステップ S 2 2 2）、未転送のデータがまだあるか否か確認し（ステップ S 2 2 3）、ある場合には、受信完了のメッセージを送信した中継装置に対して、write コマンドを送るように通信部 1 0 2 に指示する（ステップ S 2 2 4）。なお、ステップ S 2 2 4 の具体的な処理は、図 1 1 に示された処理である。

【 0 1 0 5 】

ステップ S 2 2 3 において未転送のデータがないことを確認したら、処理シーケンサ 1 0 3 は、各中継装置から受信完了のメッセージを受けるのを待ち（ステップ S 2 2 5）、データを転送した全ての中継装置から受信完了のメッセージを受けたら（ステップ S 2 2 6）、ホスト 1 0 に対して完了通知を出力するように通信部 1 0 2 に指示し（ステップ S 2 2 7）、処理を終了する。

【 0 1 0 6 】

図 1 2 は、1 ブロックのデータを 5 つに分け、3 台の中継装置 1 5 - 1, 1 5 - 2, 1 5 - 3 を使用する場合のデータ転送の例を示すタイミング図である。ストレージ 1 1 は、ホスト 1 0 からの複写コマンドに応じて、中継装置 1 5 - 1, 1 5 - 2, 1 5 - 3 のそれぞれに write コマンドを送信する。そして、受信準備完了のメッセージを送信した中継装置 1 5 - 1, 1 5 - 2, 1 5 - 3 にデータ①, ②, ③を転送する。

【 0 1 0 7 】

図 1 2 に示す例では、中継装置 1 5 - 2, 1 5 - 3 が先に完了通知を送信したので、ストレージ 1 1 は、中継装置 1 5 - 2, 1 5 - 3 に対して write コマンドを送信し、その後、データ④, ⑤を送信する。そして、全てのの中継装置 1 5 - 1, 1 5 - 2, 1 5 - 3 からの完了通知の受信を確認した時点で、ストレージ 1 1 は、ホスト 1 0 に完了通知を出力する。

【0 1 0 8】

この実施の形態では、複数の経路を用いてデータ転送を実行することができる。よって、個々の経路の使用率および中継装置の処理能力の影響を受けにくくすることができる。例えば、図 1 2 に示された例では、中継装置 1 5 - 1 への経路が輻輳していたか、または中継装置 1 5 - 1 の処理能力が低かったことになるが、他の中継装置 1 5 - 2, 1 5 - 3 を用いることによって、中継装置 1 5 - 1 への経路の輻輳または中継装置 1 5 - 1 の処理能力が低いことは、ストレージ 1 1 からのデータ転送に対して大きな影響を与えることはない。

【0 1 0 9】

ここでは、各中継装置 1 5 - 1 ~ 1 5 - n に、それぞれ異なるデータを送信する場合について説明したが、同一のデータの write コマンドを各中継装置 1 5 - 1 ~ 1 5 - n に送信するようにしてもよい。例えば、送信すべきデータとしてデータ A ~ C があるとする。ストレージ 1 1 は、データ A ~ C のいずれの write コマンドについても、各中継装置 1 5 - 1 ~ 1 5 - n に送信する。

【0 1 1 0】

この場合、ストレージ 1 1 は、各中継装置 1 5 - 1 ~ 1 5 - n に送信する同一のデータの組毎に識別子を設け、その識別子を各 write コマンドに付加する。識別子としては、送信される各データの組の順番を表す数値を用いればよい。例えば、最初にデータ A を各中継装置に送信するのであれば、各中継装置に送信する各 write コマンドに一番目を表す「1」という識別子を付加する。続いて、各中継装置にデータ B を送信するのであれば、各中継装置に送信する各 write コマンドに二番目を表す「2」という識別子を付加する。他のデータの write コマンドにも同様に識別子を付加する。

【0 1 1 1】

各中継装置 1 5 - 1 ~ 1 5 - n は、ストレージ 1 1 に対してデータの転送完了を通知するときには、write コマンドに付加された識別子も通知する。従って、ストレージ 1 1 は、同一の識別子が付加された応答を各中継装置 1 5 - 1 ~ 1 5 - n から受信する。ストレージ 1 1 は、応答に付加された識別子が初めて受信する識別子であれば、ホスト 1 0 にそのデータの転送終了を通知する。その後、同一の識別子が付加された応答を受信した場合、その応答に対しては何も処理を行わない。

【 0 1 1 2 】

また、各中継装置 1 5 - 1 ~ 1 5 - n は、識別子を付加したまま write コマンドを待機系のストレージ 1 2 に送信する。従って、ストレージ 1 2 は同一の識別子が付加された write コマンドを各中継装置 1 5 - 1 ~ 1 5 - n から受信する。ストレージ 1 2 は、write コマンドに付加された識別子が初めて受信する識別子であれば、その write コマンドに従って処理を進める。その後、同一の識別子が付加された write コマンドを受信した場合には、その write コマンドを送信した中継装置に対して、転送が完了した通知のみを行う。

【 0 1 1 3 】

このように同一データの write コマンドを各中継装置に送信すれば、複数のデータ転送経路のうち最も速度が速い経路を用いて処理を進めることができる。従って、ストレージ 1 1 がミラーリングやバックアップを行う際、中継装置からの応答の待ち時間を短くすることができる。その結果、ホスト 1 0 が次の処理を開始するまでの時間を短縮化される。

【 0 1 1 4 】

第一の実施の形態から第四の実施の形態において、中継装置 1 5 に不揮発性記憶装置を設け、中継装置 1 5 が、ストレージ 1 1 から受信したコマンドおよびデータをその不揮発性記憶装置に記録し、不揮発性記憶装置に記録したコマンドを任意のタイミングでストレージ 1 2 に発行するようにしてもよい。中継装置 1 5 に設ける不揮発性記憶装置としては、例えば、単体の磁気ディスク装置、光ディスク装置または光磁気ディスク装置の集合であるディスクアレイ装置等を使用すればよい。あるいは、バッテリバックアップされたメモリを使用してもよい。

【0115】

中継装置15は、例えば、中継装置15およびストレージ12が接続されるネットワーク14の通信量が所定のしきい値以下になった場合等に、ストレージ12に対してwrite コマンドを発行すればよい。write コマンド発行のタイミングは、不揮発性記憶装置にコマンドおよびデータを記録してから所定の時間が経過したとき、所定の時刻になったとき、あるいはストレージ12からwrite コマンドの発行を要求されたとき等であってもよい。

【0116】

このように不揮発性記憶装置にストレージ11からのコマンドおよびデータを記録すれば、コマンドやデータが喪失されることがない。そして、ストレージ11からのwrite コマンド受信と連動せずに、任意のタイミングでストレージ12にwrite コマンドを発行できる。その結果、ストレージ12を常時稼働させておかなくてもよくなる。同様に、中継装置15およびストレージ12が接続されるネットワーク14も常時接続状態を保つ必要がなくなり、ネットワーク14の運用コストも低減される。

【0117】

また、通常は、第一の実施の形態から第四の実施の形態に示したようにストレージ11からのwrite コマンド受信と連動させてストレージ12にwrite コマンドを発行し、ネットワーク14の通信量が所定のしきい値以上になったときに、コマンドやデータを不揮発性記憶装置に記録するようにしてもよい。この場合、ネットワーク14の通信量を平均化することが可能となる。また、ネットワーク14として運用コストが安価な回線を使用することができ、データ複製システム全体のコストを低減することができる。

【0118】

第一の実施の形態から第四の実施の形態において、データ転送処理手段は、処理シーケンサ103および通信部102によって実現される。中継処理手段は、中継処理部151および通信部150によって実現される。

【0119】

実施の形態5.

図 1 3 は、本発明によるデータ複製システムの第 5 の実施の形態を示すブロック図である。図 1 3 に示すデータ複製システムにおいて、ストレージ 2 0 が、ストレージ 2 0 を使用するホスト 1 0 とローカルに接続されている。ストレージ 2 0 は、ネットワーク 1 3 を介してストレージ 2 1 に接続されている。

【 0 1 2 0 】

ストレージ 2 0, 2 1 は、例えば、単体の磁気ディスク装置、光ディスク装置または光磁気ディスク装置である。ストレージ 2 0, 2 1 として、単体の磁気ディスク装置、光ディスク装置または光磁気ディスク装置の集合であるディスクアレイ装置を使用することもできる。ホスト 1 0 とストレージ 2 0 とは、SCSI、ファイバチャネル (Fibre channel)、イーサネット (登録商標) 等で接続される。

【 0 1 2 1 】

図 1 4 は、図 1 3 に示すストレージ 2 0 の構成例を示すブロック図である。なお、ストレージ 2 1 の構成も、図 1 4 に示すような構成である。図 1 4 に示すように、ストレージ 2 0 は、ストレージコントローラ 2 0 0 とストレージ本体である記憶媒体 1 0 1 とを含む。ストレージコントローラ 2 0 0 は、ホスト 1 0 および他のストレージと通信を行う通信部 2 0 4、各処理のシーケンスを管理する処理シーケンサ 2 0 1、記憶媒体 1 0 1 に対する処理命令の順序制御を行う I/O スケジューラ 1 0 4、I/O スケジューラ 1 0 4 が発行する処理命令に従って記憶媒体 1 0 1 の動作を制御する媒体処理部 1 0 5、ホスト 1 0 等から記憶媒体 1 0 1 へのデータおよび記憶媒体 1 0 1 からホスト 1 0 等へのデータを一時記憶するバッファメモリ 1 0 6、他のストレージに送るデータを冗長化する冗長化部 2 0 2、および他のストレージから送られてきた冗長化データから元のデータを復元する復元部 2 0 2 を含む。処理シーケンサ 1 0 3 は、例えば、プログラムに従って動作する CPU で実現される。

【 0 1 2 2 】

次に、実施の形態 5 のデータ複製システムの動作を、図 1 5 および図 1 6 のフローチャートを参照して説明する。図 1 5 は、ストレージコントローラ 2 0 0 における処理シーケンサ 2 0 1 の動作を示すフローチャートであり、図 1 6 は通信

部 2 0 4 の動作を示すフローチャートである。

【 0 1 2 3 】

処理シーケンサ 2 0 1 は、通信部 2 0 4 を介してホスト 1 0 から複写コマンドを受け取ると、図 1 5 に示すように、まず、複写対象のデータの範囲の先頭のブロックをデータ転送対象のブロックとして設定する（ステップ S 3 0 0）。次いで、通信部 2 0 4 に、複写コマンドで指定された複写先のストレージに対して冗長化 write コマンドを送信するように指示する（ステップ S 3 0 1）。冗長化 write コマンドは、冗長化されたデータに基づく書き込みを指示するコマンドである。冗長化 write コマンドには、ブロックサイズすなわちデータ量を示す情報とともに、データが冗長化されることを示す情報が含まれている。また、処理シーケンサ 2 0 1 は、データ転送対象のデータの読み出し要求を I O スケジューラ 1 0 4 に登録する。I O スケジューラ 1 0 4 は、読み出し要求に応じて、媒体制御部 1 0 5 に、データ転送対象のデータの読み出し指示を行う。媒体制御部 1 0 5 は、読み出し指示に従って、データ転送対象のデータを記憶媒体 1 0 1 からバッファメモリ 1 0 6 に出力させる（ステップ S 3 0 2）。

【 0 1 2 4 】

そして、処理シーケンサ 2 0 1 は、通信部 2 0 4 を介して入力されるストレージ 2 1 からの受信準備完了のメッセージと、媒体制御部 1 0 5 からの読み出し完了通知との双方を待ち（ステップ S 3 0 3）。双方が入力されたら、バッファメモリ 1 0 6 に格納されたデータの冗長化を冗長化部 2 0 2 に行わせる（ステップ S 3 0 4）。冗長化は、元のデータ（以下、元データという。）から冗長データを作成することによって行う。以下の説明において、元データおよび新たに作成された冗長データの集合を冗長化されたデータ群と記す。処理シーケンサ 2 0 1 は、冗長化されたデータ群をストレージ 2 1 に転送するように通信部 2 0 4 に指示する。通信部 2 0 4 は、この指示に応じて、冗長化されたデータ群をストレージ 2 1 に送信する（ステップ S 3 0 5）。

【 0 1 2 5 】

ただし、通信部 2 0 4 は、元データと冗長データとを別々のデータのまとまりとしてストレージ 2 1 に送信する。この「データのまとまり」とは、各種データ

転送プロトコルの最小のデータ送信単位のことである。「データのまとまり」の具体例としては、例えば、TCPやUDPにおけるセグメント、インターネットプロトコルにおけるパケット、あるいはファイバチャネルプロトコルやイーサネット（登録商標）の物理層におけるフレーム等がある。従って、例えばインターネットプロトコルに従って送信を行う場合、通信部204は、元データと冗長データとを別々のパケットで送信する。なお、一般に、パケット等のデータ送信単位の内部には、伝送中に生じたエラー（データ化け）を検出するためのエラー検出データが付加される。ステップS304で作成される冗長データは、独立したデータのまとまりとして送信されるデータであり、パケット等に付加されるエラー検出データとは異なるデータである。

【0126】

また、後述するように、元データを分割し、分割された元データに基づいて冗長データを作成する。ステップS305において、通信部305は、分割された元データをそれぞれ個々のデータ送信単位として送信する。例えば、元データをN個に分割し、N個のデータからm個の冗長データを作成したとする。この場合、通信部は元データおよび冗長データをそれぞれN個、m個のパケット（あるいはセグメント、フレーム等）で送信する。通信部204が元データと冗長データとをそれぞれ別々のデータのまとまりとして送信したならば、通信部204を介して入力されるストレージ21からの受信完了のメッセージを待つ（ステップS306）。

【0127】

通信部204は、冗長化されたデータ群（元データと冗長データ）を全てストレージ21に送信し、ストレージ21から受信完了のメッセージを受けたら、受信完了のメッセージを処理シーケンサ201を出力する。処理シーケンサ201は、受信完了のメッセージを入力すると、複写対象の全てのデータの中継装置15への転送が完了したか否か確認する（ステップS307）。完了していなければ、複写対象のデータの範囲の次のブロックをデータ転送対象のブロックとして設定し（ステップS309）、ステップS301に戻る。

【0128】

複写対象の全てのデータの転送が完了していれば、処理シーケンサ 2 0 1 は、ホスト 1 0 に対して完了通知を出力するように通信部 2 0 4 に指示し（ステップ S 3 0 8）、処理を終了する。

【 0 1 2 9 】

なお、1 ブロックのデータ量は、あらかじめシステムに設定されている量である。また、処理シーケンサ 2 0 1 は、1 ブロックのデータ量を、転送先のストレージ 2 1 のバッファメモリ 1 0 6 の容量に応じて変化させるようにしてもよい。

【 0 1 3 0 】

通信部 2 0 4 は、外部から入力されたデータが、コマンド、受信準備完了のメッセージ、受信完了のメッセージ等の制御系メッセージであった場合には、入力されたデータを処理シーケンサ 2 0 1 に渡す。また、ストレージ 2 0 に書き込まれるべきデータであった場合には、データを格納すべき場所を処理シーケンサ 2 0 1 に問い合わせる。そして、処理シーケンサ 2 0 1 から指定されたバッファメモリ 1 0 6 中の領域にデータを格納するか、または、復元部 2 0 3 にデータを復元させた後、指定されたバッファメモリ 1 0 6 中の領域にデータを格納する。

【 0 1 3 1 】

また、通信部 2 0 4 は、処理シーケンサ 2 0 1 から指定されたコマンドまたは完了通知を、指定されたストレージまたはホストに送信する処理も行う。さらに、処理シーケンサ 2 0 1 から指定されたバッファメモリ 1 0 6 中のデータを、指定されたストレージまたはホストに送信する処理も行う。また、冗長化部 2 0 2 から冗長化されたデータを受け取り、受け取ったデータを、指定されたストレージまたはホスト 1 0 に送信する処理も行う。

【 0 1 3 2 】

なお、通信部 2 0 4 は、他のストレージにコマンドを送信する場合に、各コマンドを区別するためのコマンド識別子をコマンドに付加する。コマンド識別子の例として、コマンドを送信する毎に 1 加算した数値である発行番号（後述する発行 ID）がある。また、冗長化されたデータ群を送信する際に、その送信に関連したコマンドのコマンド識別子を付加する。例えば、冗長化 write コマンドを送信するときに、その冗長化 write コマンドを識別するためのコマンド識別子を付

加する。そして、その冗長化write コマンドに対応する冗長化されたデータ群（元データおよび冗長データ）にも、冗長化write コマンドと同じコマンド識別子を付加する。

【 0 1 3 3 】

さらに、通信部 2 0 4 は、各コマンドやデータに、送信元となるストレージの識別子も付加する。

【 0 1 3 4 】

次に、冗長化部 2 0 2 の動作について説明する。冗長化部 2 0 2 は、指定されたデータをバッファ 1 0 6 から取得し、規定された冗長化方法を用いて、冗長化されたデータ群を作成する。冗長化部 2 0 2 は、指定された元データを N 個のデータに分割する。そして、 N 個に分割された元データから m 個の冗長データを作成する。 N 、 m は自然数である。この $N + m$ 個のデータ群が、冗長化されたデータ群となる。冗長化されたデータ群には、復元時に使用するための識別情報が付加される。

【 0 1 3 5 】

以下、冗長化の具体例について説明する。冗長化部 2 0 2 は、例えば、元データをその先頭から N 等分するなどの方法により、元データを N 個に分割する。この方法を採用した場合、先頭から何番目のデータであるのかを示す番号を識別情報として使用する。冗長化部 2 0 2 は、元データを N 個のデータに分割したならば、RAID 3 および RAID 5 の装置等と同様にパリティ演算によってパリティデータを作成し、そのパリティデータを冗長データとする。すなわち、 N 個に分割された各データおよび冗長データにおいて、対応する各ビットの値の総和が必ず奇数（または必ず偶数）になるように冗長データを作成する。この場合、1 個の冗長データを作成すればよい。このように冗長ビットを作成すれば、分割後の N 個の元データのうち、一つが欠落しても、欠落したデータを復元できる。

【 0 1 3 6 】

冗長化の方法は上記のパリティ演算のみに限定されず、冗長データの数も 1 個とは限らない。例えば、ダブルパリティ演算によって冗長データを作成してもよい。また、元データに基づく ECC (Error Correcting Code: 誤り訂正符号)

を冗長データとして用いてもよい。このように様々な冗長化の方法があるが、以下の説明では、パリティ演算によって1個の冗長データを作成した場合を例に説明する。

【0137】

次に、データの転送先であるストレージ21の動作を説明する。ストレージ20からストレージ21のストレージコントローラ200における通信部204に冗長化write コマンドが送られる。通信部204は、冗長化write コマンドを受信すると、処理シーケンサ201に冗長化write コマンドを渡し、冗長化write コマンドにもとづく処理（冗長化write 処理）の開始を指示する。

【0138】

図16は、ストレージ21における処理シーケンサ201の動作を示すフローチャートである。処理シーケンサ201は、複数の冗長化write 処理を並行して行うことが可能であり、コマンドを発行処理したストレージの識別子と、コマンドに付加されたコマンド識別子を用いて各処理を識別する。

【0139】

冗長化write 処理では、処理シーケンサ201は、まず、送られてくるデータを記憶できる領域をバッファ106中に確保する（ステップS320）。次いで、通信部204に冗長化write コマンドを送信したストレージ20に向けて受信準備の完了通知を送信させる（ステップS321）。そして、ストレージ20からデータが届くのを待ち（ステップS322）、データが到着したことが通信部204から通知されたら、データ転送元のストレージ20の識別子とデータに付加されているコマンド識別子とを指定し、データを復元部203に送るように通信部204に指示する（ステップS323）。

【0140】

次いで、処理シーケンサ201は、冗長化write コマンドに対応して送られてきたデータの個数を判断する（ステップS324）。各データには、冗長化write コマンドに対応したコマンド識別子が付加されているので、処理シーケンサ201は、受信した各データが冗長化write コマンドに対応して送られてきたデータか否かを確認することができる。データの個数が所定の個数（n個）未満の場合

合にはステップ S 3 2 2 に移行し、所定の個数（ n 個）である場合にはステップ S 3 2 5 に移行する（ステップ S 3 2 4）。この所定の個数は、元データを復元することができるデータの個数であり、冗長化の方式によって異なる。例えば、 N 個に分割された元データからパリティ演算によって 1 個の冗長データが作成された場合、ストレージ 2 0 からは $N + 1$ 個のデータが送られてくる。この場合、 N 個のデータを受信すれば、元データを復元することができる。従って、ステップ S 3 2 4 では受信したデータの数 N 個未満か否かを判断すればよい。

【 0 1 4 1 】

ステップ S 3 2 5 において、処理シーケンサ 2 0 1 は、データ転送元のストレージ 2 0 の識別子およびコマンド識別子と、ステップ S 3 2 0 で確保したバッファ 1 0 6 中の領域とを指定し、データの復元を復元部 2 0 3 に指示する。そして、データ転送元のストレージ 2 0 に応答を送信するように通信部 2 0 4 に指示する（ステップ S 3 2 6）。また、I/O スケジューラ 1 0 4 に対して、ステップ S 3 2 0 で確保したバッファ 1 0 6 中の領域中のデータを、冗長化 write コマンドで指定された媒体中の領域に書き込むように指示を出し（ステップ S 3 2 7）、処理を終了する。I/O スケジューラ 1 0 4 にこの指示が出された場合の I/O スケジューラ 1 0 4 および媒体制御部 1 0 5 の動作は、第 1 の実施の形態で説明したステップ S 1 0 2 における動作と同様である。

【 0 1 4 2 】

なお、処理シーケンサ 2 0 1 は、ステップ S 3 2 5 の処理開始以降に通信部 2 0 4 に到着した冗長化 write コマンドに関連したデータについては、通信部 2 0 4 にそのデータの破棄を指示する。例えば、 N 個のデータから元データを復元できる場合、 $N + 1$ 個目に到着したデータは必要がない。必要なデータが到着した後にストレージ 2 1 に到着したデータは破棄する。

【 0 1 4 3 】

次に、復元部 2 0 3 の処理について説明する。復元部 2 0 3 は、バッファメモリを有し、データ転送元のストレージ 2 0 の識別子およびコマンド識別子とともに渡されたデータをバッファメモリに蓄える。データ転送元のストレージ 2 0 の識別子およびコマンド識別子と、バッファメモリ 1 0 6 中の復元したデータを格

納する領域とが指定され、復元が指示されると、復元部 2 0 3 は、内部のバッファから、データ転送元のストレージ 2 0 の識別子およびコマンド識別子に該当するデータを集める。そして、各データに付加された識別情報を元に、データ群から冗長化前のデータを復元する。

【 0 1 4 4 】

例えば、RAID 3 および RAID 5 の装置等のように、N 個に分割された元データからパリティ演算によって 1 個の冗長データを作成して冗長化を行っていたとする。そして、分割された N 個の元データのうち k 番目のデータがストレージ 2 1 に到着せず、他の N - 1 個のデータと 1 個の冗長データ（合計 N 個のデータ）が到着したとする。この場合、k 番目のデータと到着した N 個のデータにおいて、対応する各ビットの値の総和が必ず奇数（または必ず偶数）になるようにすることによって、k 番目のデータを作成することができる。そして、作成した k 番目のデータと、ストレージ 2 1 に到着した他のデータとによって、分割前の元データを復元すればよい。分割された各データには、識別情報（例えば、先頭から何番目のデータであるのかを示す番号）が付加されているので、分割された状態から分割前の元データに復元させることができる。また、N 個に分割された元データが全てストレージ 2 1 に到着し、冗長データが到着しなかった場合には、到着した N 個のデータから分割前の元データを復元すればよい。

【 0 1 4 5 】

復元部 2 0 3 は、指定されたバッファメモリ 1 0 6 中の領域に復元したデータを格納する。なお、復元部 2 0 3 の内部のバッファにおける復元処理で使用されていたデータを格納していた領域は、復元処理終了後他のデータの格納に再利用される。

【 0 1 4 6 】

既述のように、冗長化の方法は、パリティ演算によって 1 個の冗長データを作成する方法に限られない。転送先のストレージ 2 1 では、冗長化の方法に対応した方法で復元を行えるように制御すればよい。例えば、ECC を冗長データとしたならば、ECC に対応した方法で復元を行うように制御すればよい。

【 0 1 4 7 】

また、ここでは、元データをN個に分割して冗長データを作成する場合を例示に説明したが、ストレージ20の冗長化部202は、ステップS304において、元データを分割せずに元データと同一のデータを複製し、その複製したデータを冗長データとしてもよい。そして、ステップS305では、元データを一つのデータのまとまり（例えば、一つの packets、セグメントあるいはフレーム等）として送信し、元データの複製である冗長データも一つのデータのまとまりとして送信してもよい。元データと冗長データとは同一のデータであり、また、分割されていないので、転送先ストレージ21では、いずれか一方を受信したときに、受信したデータをそのまま記憶媒体101に書き込んでよい。また、遅れて受信したデータは破棄してよい。従って、ストレージ21の処理シーケンサ201は、一回目にステップS324に移行したならば、すぐに次のステップS325に移行してよい。さらに、ステップS325において復元部203が復元処理を行うことなく、次のステップS326に移行してよい。

【0148】

このように元データと冗長データとを同一のデータとして、別々の二つのデータのまとまりとして送信すれば、一方のデータが送信過程で失われてもストレージ21は、送信元が送信しようとした元データをそのまま受信することができる。

【0149】

また、ここでは、元データを分割せずに元データと同一のデータを複製し、その複製したデータを冗長データとする場合を示した。ストレージ20の冗長化部202は、冗長データを作成するときに、元データをN個に分割し、分割したN個のデータについてそれぞれ同一のデータを複製してもよい。この場合、分割後のN個のデータの複製であるN個のデータが冗長データとなる。そして、ステップS305では、元データをN個のデータのまとまり（例えば、N個の packets、セグメントあるいはフレーム等）として送信し、冗長データもN個のデータのまとまりとして送信してもよい。この場合、ストレージ21には $2 \times N$ 個のデータが送信される。ストレージ21は、この $2 \times N$ 個のデータのうち、元データを復元することができる一組のデータ（分割後の第一番目から第N番目までの各デ

ータ) について受信したならば、そのデータに基づいて元データを復元する。すなわち、ストレージ 2 1 の処理シーケンサ 2 0 1 は、ステップ S 3 2 4 において、分割後の第一番目のデータから第 N 番目のデータまでを全て受信したか否かを確認すればよい。そして、全て受信していればステップ S 3 2 5 に移行し、まだ受信していないデータがあればステップ S 3 2 2 に戻ればよい。

【 0 1 5 0 】

このように複製データを冗長データとして送信すれば、 $2 \times N$ 個のデータの一部が送信過程で廃棄されても、ストレージ 2 1 が分割後の第一番目のデータから第 N 番目のデータまでを全て受信した時点でデータを復元することができる。なお、ストレージ 2 0 の冗長化部 2 0 2 は、元データを分割してからその分割後の各データの複製を作成してもよいし、また、分割前の元データと同一のデータを複製し、元データおよび複製データをそれぞれ分割してもよい。

【 0 1 5 1 】

第 5 の実施の形態によれば、冗長データを作成し、元データと冗長データとを別々に送信する。従って、データの一部が送信過程で失われてしまったとしても、元データを復元することができる。あるいは、送信元が送信しようとした元データをそのまま受信することができる。この結果、元データの一部が送信過程で失われた場合、転送元のストレージ 2 0 に元データが届いていないことを通知して、ストレージ 2 0 に再度送信させる必要がなくなり、データの転送時間を短縮することができる。

【 0 1 5 2 】

図 1 3 に示すデータ複製システムでは、冗長化されたデータ群に含まれる元データおよび冗長データは、いずれもネットワーク 1 3 を介して送受信される。元データと冗長データとを別々のネットワークを介して送信する構成であってもよい。図 1 7 は、元データと冗長データとを別々のネットワークを介して送信するデータ複製システムの構成例を示すブロック図である。図 1 7 に示す構成例において、ストレージ 2 0 は、ネットワーク 1 3 を介してストレージ 2 1 に接続され、またネットワーク 1 4 を介してもストレージ 2 1 に接続されている。ストレージ 2 0 とストレージ 2 1 とが二つのネットワーク 1 3, 1 4 を介して接続されて

いる点以外は、図 1 3 に示す構成と同様である。また、図 1 7 に示すストレージ 2 0、2 1 の構成は、図 1 4 に示す構成と同様である。ただし、通信部 2 0 4 は、二つのネットワーク 1 3、1 4 に接続される。

【0 1 5 3】

また、図 1 7 に示すストレージ 2 0 の動作は、図 1 3 に示すストレージ 2 0 の動作と同様である。ただし、ステップ S 3 0 5（図 1 5 参照）において、通信部 2 0 4 が冗長化されたデータ群をストレージ 2 1 に送信する場合、通信部 2 0 4 は、元データと冗長データをそれぞれ別々のネットワークを介して送信する。例えば、元データを一つのデータ転送単位として（例えば一つの packets 等として）、ネットワーク 1 3 a を介してストレージ 2 1 に送信し、元データの複製である冗長データを、ネットワーク 1 4 を介してストレージ 2 1 に送信する。

【0 1 5 4】

このような構成によれば、ネットワーク 1 3、1 4 のいずれか一方で障害が発生したり、いずれか一方のネットワークの転送速度が遅かったとしても、もう一方のネットワークを介して元データまたは元データと同一の冗長データを送信することができる。従って、元データの複製を冗長データとする場合に、耐障害性をより高めることができる。なお、元データを N 個に分割して冗長データを作成する場合であっても、分割した元データと冗長データとをそれぞれ別のネットワークを介して送信するようにしてもよい。

【0 1 5 5】

実施の形態 6.

第 5 の実施の形態では、ストレージ 2 0 のデータのバックアップが実現されたが、ミラーリングによってストレージ 2 0 のデータをストレージ 2 1 に転送するようにしてもよい。図 1 8 は、第 6 の実施の形態、すなわちミラーリングを行う場合のストレージ 2 0 のストレージコントローラ 2 0 0 における処理シーケンス 2 0 1 の動作を示すフローチャートである。なお、データ複製システムの構成およびストレージ 2 0、2 1 の構成は第 5 の実施の形態の場合と同じである（図 1 3、図 1 4 参照）。また、ストレージ 2 0 において、通信部 2 0 4、I/O スケジューラ 1 0 4 および媒体制御部 1 0 5 の動作は、第 5 の実施の形態のそれらの動

作と同じである。

【 0 1 5 6 】

ストレージ 2 0 は、ホスト 1 0 から write コマンドを受信したときにミラーリングを開始する。ストレージ 2 0 の通信部 2 0 4 がホスト 1 0 から write コマンドを受信すると、通信部 2 0 4 はその write コマンドを処理シーケンサ 2 0 4 に渡す。すると、処理シーケンサ 2 0 1 は、図 1 8 に示すように、ホスト 1 0 から受け取るデータを格納するのに必要な領域をバッファメモリ 1 0 6 に確保する（ステップ S 3 4 0）。また、準備完了の通知をホスト 1 0 に送るように通信部 2 0 4 に指示する（ステップ S 3 4 1）。通信部 2 0 4 は、指示に応じて、準備完了の通知をホスト 1 0 に送る。

【 0 1 5 7 】

そして、処理シーケンサ 2 0 1 は、ホスト 1 0 からデータが到着するのを待ち（ステップ S 3 4 2）、データが届いて通信部 2 0 4 からデータを格納すべきバッファメモリ 1 0 6 の領域の問い合わせを受けると、ステップ S 3 4 0 で確保した領域を通信部 2 0 4 に知らせる（ステップ S 3 4 3）。次いで、冗長化 write コマンドをストレージ 2 1 に送るように通信部 2 0 4 に指示する（ステップ S 3 4 4）。通信部 2 0 4 は、指示に応じて、ストレージ 2 1 に冗長化 write コマンドを送信する。

【 0 1 5 8 】

次いで、処理シーケンサ 2 0 1 は、ホスト 1 0 からのデータのバッファメモリ 1 0 6 への格納の完了を待ち（ステップ S 3 4 5）、全てのデータがバッファメモリ 1 0 6 に格納されたことが通信部 2 0 4 から通知されると、処理シーケンサ 2 0 1 は、I/Oスケジューラ 1 0 4 に対して、処理の種類（この場合には書き込み）と、処理の識別 ID と、処理の対象となる記憶媒体 1 0 1 中の領域を示す情報と、処理の対象となるバッファメモリ 1 0 6 の領域を示す情報とを登録する。I/Oスケジューラ 1 0 4 は、書き込み要求に応じて、媒体制御部 1 0 5 に、書き込み対象のデータの書き込み指示を行う。媒体制御部 1 0 5 は、登録内容に応じてバッファメモリ 1 0 6 から記憶媒体 1 0 1 へのデータの書き込み処理を行う（ステップ S 3 4 6）。

【0159】

そして、処理シーケンサ201は、ストレージ21から受信準備完了のメッセージ（ステップS344において送った冗長化write コマンドに対する応答）が送信されるのを待つ（ステップS347）。ストレージ21からの受信準備完了のメッセージを受信したことが通信部204から通知されると、バッファメモリ106に格納されたデータを冗長化するように冗長化部202に指示する。冗長化部202は、指示に応じて、第5の実施の形態の場合と同様に冗長化を行う（ステップS348）。ステップS348では、元データをN個に分割し、分割後のデータから冗長データを作成してもよい。あるいは、元データを分割せず、元データを複製したデータを冗長データとしてもよい。また、元データを分割した各データと同一のデータを複製し、複製した各データを冗長データとしてもよい。元データと同一のデータを複製し、元データおよび複製データをそれぞれ分割してもよい。

【0160】

続いて、処理シーケンサ201は、通信部204に、冗長化部202によって冗長化されたデータ群をストレージ21に向けて送信するように指示する（ステップS349）。通信部204は、この指示に応じて、冗長化されたデータ群（分割後の元データおよび冗長データ）をストレージ21に送信する。その後、ストレージ21から受信完了のメッセージが送信されるのと、媒体制御部105からの書き込み完了通知とを待ち（ステップS350）、ストレージ21からの受信完了のメッセージを受信したことが通信部204から通知され、かつ、媒体制御部105からの書き込み完了通知を受けると、ホスト10に完了通知し（ステップS351）、処理を終了する。なお、ストレージ21が、ストレージ20から冗長化されたデータ群を受信するときの動作は、実施の形態5と同様である。

【0161】

図17に示す場合と同様に、ストレージ20とストレージ21とがネットワーク13、14によって接続されていてもよい。そして、ステップS349において、通信部204が冗長化されたデータ群を送信するときには、元データと冗長データとをそれぞれ別々のネットワークを介して送信するようにしてもよい。

【 0 1 6 2 】

第 1 の実施の形態から第 4 の実施の形態のデータ複製システムにおいて、ストレージと中継装置とがデータを送受信する際に、第 5 の実施の形態または第 6 の実施の形態に示したように冗長化されたデータ群を送受信するようにしてもよい。その場合、第 1 の実施の形態から第 4 の実施の形態においても、第 5 の実施の形態または第 6 の実施の形態と同様の効果が得られる。

【 0 1 6 3 】

第 5 の実施の形態および第 6 の実施の形態において、データ転送処理手段は、処理シーケンサ 2 0 1 および通信部 2 0 4 によって実現される。冗長化手段 2 0 2 は、冗長株 2 0 2 によって実現される。復元手段は、復元部 2 0 3 によって実現される。格納処理手段は、処理シーケンサ 2 0 1、I/O スケジューラ 1 0 4 および媒体制御部 1 0 5 によって実現される。

【 0 1 6 4 】

実施の形態 7.

図 1 9 は、本発明によるデータ複製システムの第 7 の実施の形態を示すブロック図である。図 1 9 に示すデータ複製システムにおいて、ストレージ 3 0 1 が、ストレージ 3 0 1 を使用するホスト（上位装置）3 0 0 とローカルに接続されている。また、ストレージ 3 0 2 が、ストレージ 3 0 2 を使用するホスト 3 0 3 とローカルに接続されている。ストレージ 3 0 1 は、ネットワーク 1 3 を介してストレージ 3 0 2 に接続されている。また、ホスト 3 0 0 とホスト 3 0 3 とは通信可能に接続されている。ホスト 3 0 0 とホスト 3 0 3 とは、専用回線によって接続されていることが好ましいが、専用回線以外のネットワーク（例えばインターネット等）によって接続されていてもよい。

【 0 1 6 5 】

ストレージ 3 0 1、3 0 2 は、例えば、単体の磁気ディスク装置、光ディスク装置または光磁気ディスク装置である。ストレージ 3 0 1、3 0 2 として、単体の磁気ディスク装置、光ディスク装置または光磁気ディスク装置の集合であるディスクアレイ装置を使用することもできる。ホスト 3 0 0、3 0 3 とストレージ 3 0 1、3 0 2 とは、SCSI、ファイバチャネル（Fibre channel）、イーサ

ネット（登録商標）等で接続される。なお、図 1 9 に示すシステムにおいて、ホスト 3 0 0 が、システムに障害が発生していないときに稼働する正常系ホストであり、ホスト 3 0 3 が、ホスト 3 0 0 において障害が発生したときに稼働する待機系ホストであるとする。

【 0 1 6 6 】

ホスト 3 0 0, 3 0 3 は、それぞれホスト 3 0 0, 3 0 3 自身が保持するアプリケーションプログラム（以下、アプリケーションと記す。）に従って処理を行う。このアプリケーションは、ストレージ 3 0 1, 3 0 2 のデータを処理対象とする。例えば、ストレージ 3 0 1, 3 0 2 に銀行の顧客の預金額等のデータを記憶する場合には、ホスト 3 0 0, 3 0 3 は、預金データ管理アプリケーションに従って、ストレージ 3 0 1, 3 0 2 内のデータの更新等を行う。実際に動作を行うのはホストであるが、以下、アプリケーションの動作として説明する。

【 0 1 6 7 】

本実施の形態では、正常系のストレージ 3 0 1 内のデータが遠隔地に存在する待機系のストレージ 3 0 3 にミラーリングされる。ただし、正常系ホスト 3 0 0 がストレージ 3 0 1 にデータを書き込んだ時点では、待機系ストレージ 3 0 3 の記憶媒体にそのデータを書き込ませるのではなく、待機系ストレージ 3 0 3 が備える同期用バッファメモリにそのデータを保持させる。そして、ホスト 3 0 0 が指定するタイミングで、待機系ストレージ 3 0 3 の同期用バッファメモリ内のデータを所定の記憶媒体に書き込ませる。例えば、ホストが処理 X を開始するときには、ストレージにデータ A, B が書き込まれていなければならないとする。この場合、正常系のホスト 3 0 0 がストレージ 3 0 1 にデータ A を書き込んだ時点では、待機系ストレージ 3 0 2 にはデータ A を同期用バッファメモリに保持させるだけで、待機系ストレージ 3 0 2 の記憶媒体への書き込みは行わせない。ホスト 3 0 0 がストレージ 3 0 1 にデータ B を書き込んだ後に、ホスト 3 0 0 は待機系ストレージ 3 0 2 の記憶媒体への書き込みタイミングを指定し、そのタイミングで待機系ストレージ 3 0 2 は同期用バッファメモリ内のデータ A, B を記憶媒体に書き込む。

【 0 1 6 8 】

そのデータの状態であればアプリケーションがそのまま動作を再開することができるタイミングを、再開可能ポイントと記す。すなわち、再開可能ポイントとは、書き込まれたデータの状態がアプリケーションによる処理を再開することができる状態になっているタイミングのことである。上記の例では、正常系のストレージ 3 0 1 にデータ A, B が書き込まれてから次のデータが書き込まれるまでの間、処理 X を開始することができる再開可能ポイントとなる。

【 0 1 6 9 】

図 2 0 は、ホスト 3 0 0 の構成例を示すブロック図である。ホスト 3 0 0 において、単数あるいは複数のアプリケーションが動作する。ここでは、2 つのアプリケーション 3 1 0 a, 3 1 0 b を例示する。アプリケーション 3 1 0 a, 3 1 0 b は、I/O 管理部 3 1 1 を用いて、ストレージ 3 0 1 中のデータにアクセスする。また、I/O 管理部 3 1 1 は、ストレージ 3 0 1 に再開可能ポイントを通知するための再開可能ポイント通知部 3 1 2 を有する。アプリケーション 3 1 0 a, 3 1 0 b に従い、再開可能ポイント通知部 3 1 2 は、再開可能ポイントにおいて、ストレージ 3 0 1 に再開可能ポイントであることを知らせる処理（再開可能ポイント通知処理）を行う。また、ホスト 3 0 0 の状態を監視するホスト監視部 3 1 3 が備えられている。なお、ホスト 3 0 3 の構成は、ホスト 3 0 0 の構成と同じである。

【 0 1 7 0 】

本実施の形態におけるアプリケーション 3 1 0 a, 3 1 0 b は、再開機能を有するアプリケーションである。すなわち、ストレージの記憶媒体 1 0 1 のデータ記録状態が所定の状態になっていれば処理を再開できる機能を実現するアプリケーションである。

【 0 1 7 1 】

図 2 1 は、図 1 9 に示すストレージ 3 0 1 の構成例を示すブロック図である。なお、ストレージ 3 0 2 の構成も、図 2 1 に示すような構成である。図 2 1 に示すように、ストレージ 3 0 1 は、ストレージコントローラ 3 2 0 とストレージ本体である記憶媒体 1 0 1 とを含む。ストレージコントローラ 3 2 0 は、ホスト 3 0 0 および他のストレージと通信を行う通信部 3 2 2、各処理のシーケンスを管

理する処理シーケンサ 3 2 1、記憶媒体 1 0 1 に対する処理命令の順序制御を行う I O スケジューラ 1 0 4、I O スケジューラ 1 0 4 が発行する処理命令に従って記憶媒体 1 0 1 の動作を制御する媒体処理部 1 0 5、ホスト 3 0 0 から記憶媒体 1 0 1 へのデータおよび記憶媒体 1 0 1 からホスト 3 0 0 へのデータを一時記憶するバッファメモリ 1 0 6、および他のストレージから送られてきたデータを一時保存する同期用バッファメモリ 3 2 2 を含む。処理シーケンサ 3 2 1 は、例えば、プログラムに従って動作する CPU で実現される。I O スケジューラ 1 0 4 および媒体制御部 1 0 5 の動作は、第 1 の実施の形態における I O スケジューラ 1 0 4 および媒体制御部 1 0 5 の動作と同様である。

【 0 1 7 2 】

なお、同期用バッファメモリ 3 2 2 として半導体メモリが使用される場合もあるし、磁気ディスク装置、光ディスク装置または光磁気ディスク装置等のより大容量の記憶装置が使用される場合もある。

【 0 1 7 3 】

アプリケーション 3 1 0 a、3 1 0 b がストレージ 3 0 1 からデータを読み出す際、ホスト 3 0 0 は、ストレージ 3 0 1 にデータの読み出しを要求する。ストレージ 3 0 1 の処理シーケンサ 3 2 1 は、I O スケジューラ 1 0 4 等によって、要求されたデータを記憶媒体 1 0 1 からバッファメモリ 1 0 6 にコピーする。そして、そのデータをホスト 3 0 0 に送信する。

【 0 1 7 4 】

次に、ホスト 3 0 0 がストレージに 3 0 1 にデータを書き込むときの動作の概要について説明する。ホスト 3 0 0 は、write コマンドをストレージ 3 0 1 に出力することによって、ストレージ 3 0 1 にデータを書き込ませる。また、一回または複数回 write コマンドを出力した後に再開可能ポイントになったならば、再開可能ポイント通知コマンドをストレージ 3 0 1 に出力することによって、再開可能ポイントになったことを通知する。

【 0 1 7 5 】

ストレージ 3 0 1 は、write コマンドを受信した場合、その write コマンドに従って、記憶媒体 1 0 1 にデータを書き込む。また、ストレージ 3 0 2 に、遅延

write コマンドを出力して、そのデータを待機系ストレージ 3 0 2 の同期用バッファメモリ 3 2 3 に保持させる。遅延 write コマンドは、記憶媒体 1 0 1 に書き込むべきデータを同期用バッファメモリ 3 2 3 に保持させ、後述する遅延データ反映コマンドが届いたときに記憶媒体 1 0 1 に書き込むことを指示するコマンドである。

【 0 1 7 6 】

ストレージ 3 0 1 は、再開可能ポイント通知コマンドを受信した場合、遅延データ反映コマンド（遅延書き込み実行要求）をストレージ 3 0 2 に送信する。遅延データ反映コマンドは、同期用バッファメモリ内に記憶しているデータを記憶媒体 1 0 1 に書き込むことを指示するコマンドである。待機系のストレージ 3 0 2 は、遅延データ反映コマンドを受信したときに、同期用バッファメモリ 3 2 3 に記憶したデータを記憶媒体 1 0 1 に書き込む。このような動作によって、ストレージ 3 0 2 の記憶媒体 1 0 1 が常に処理を開始できる状態に保つ。

【 0 1 7 7 】

ストレージ 3 0 1 は、遅延 write コマンドおよび遅延データ反映コマンドに同期 ID および発行 ID を付加して送信する。図 2 2 は、同期 ID および発行 ID の例を示す説明図である。ストレージ 3 0 1 は、ホスト 3 0 0 から再開可能ポイント通知コマンドを受信する度に同期 ID を更新する。従って、ある再開可能ポイントから次の再開可能ポイントまでの間に出力した遅延 write コマンドには、同一の同期 ID が付加される。ただし、ストレージ 3 0 1 は、再開可能ポイント通知コマンドを受信したときに発行する遅延データ反映コマンドに、更新する直前の同期 ID を付加する。その後に発行する遅延 write コマンドに更新後の同期 ID を付加する。図 2 2 に示す例では、遅延 write コマンドに更新後の同期 ID である「2 4」や「2 5」が付加され、その直近に出力された遅延データ反映コマンドに更新前の同期 ID である「2 3」や「2 4」が付加されている。

【 0 1 7 8 】

また、ストレージ 3 0 1 は、遅延 write コマンドまたは遅延データ反映コマンドを発行する度に（すなわち、遅延 write コマンドまたは遅延データ反映コマンドを作成して送信する度に）発行 ID を更新する。図 2 2 に示す例では、遅延 wr

ite コマンドまたは遅延データ反映コマンドの発行IDが1増加する場合の例を示している。

【0179】

ストレージ301は、ホスト300からのwrite コマンドを受信すると、ストレージ302に遅延write コマンドを送信し、その後、遅延write コマンドに対応するデータを送信する。このデータの送信順序は、遅延write コマンドの発行順序と同一でなくてもよい。例えば、ストレージ301が、データp, qを書き込ませるために遅延write コマンドを発行し、その後データrを書き込ませるために次の遅延write コマンドを発行するものとする。この場合、ストレージ301が各データを送信する順番は、p, q, rという順番に限られない。データp, qの送信完了前にデータrの送信を開始してもよい。ただし、遅延データ反映コマンドをストレージ302に送信するときには、それまでに発行した遅延write コマンドのデータの送信を完了させてから遅延データ反映コマンドを送信する。

【0180】

なお、ストレージ301は、遅延write コマンドに対して、遅延write コマンドで指定されたデータを書き込むべき場所（オフセットアドレス、セクター番号、ブロック番号等）やデータのサイズも付加する。

【0181】

次に、アプリケーション310a, 310bが、ストレージ301にデータを書き込むときの動作を説明する。図23は、ストレージコントローラ320における処理シーケンサ321の動作を示すフローチャートである。

【0182】

アプリケーション310a, 310bは、ストレージ301にデータを書き込むときに、ストレージ301のストレージコントローラ320における通信部322に対してwrite コマンドを出力する。通信部322は、write コマンドを受け取ると、write コマンドを処理シーケンサ321に渡し、write 処理を開始することを処理シーケンサ321に指示する。

【0183】

処理シーケンサ 3 2 1 は、write コマンドを受け取ると、図 2 3 に示すように、ホスト 3 0 0 から送られてくるデータを格納するのに必要な領域をバッファメモリ 1 0 6 に確保する（ステップ S 4 0 0）。そして、準備完了の通知をホスト 3 0 0 に送るように通信部 3 2 2 に指示する（ステップ S 4 0 1）。通信部 3 2 2 は、指示に応じて、ホスト 3 0 0 に準備完了の通知を送信する。

【 0 1 8 4 】

次いで、ホスト 3 0 0 からデータが届くのを待ち（ステップ S 4 0 2）、データが届いて通信部 3 2 2 からデータを格納すべきバッファメモリ 1 0 6 の領域の問い合わせを受けると、ステップ S 4 0 0 で確保した領域を通信部 3 2 2 に知らせる（ステップ S 4 0 3）。また、通信部 3 2 2 にストレージ 3 0 2 に対して同期 I D を指定して遅延 write コマンドを送信するよう指示する（ステップ S 4 0 4）。通信部 3 2 2 は、指示に応じて、ストレージ 3 0 2 に遅延 write コマンドを送信する。既に説明したように、遅延 write コマンドには、同期 I D および発行 I D が付加される。ある再開可能ポイントから次の再開可能ポイントまでの間に出力する各遅延 write コマンドには、同一の同期 I D を付加する。また、遅延 write コマンドまたは遅延データ反映コマンドを出力する度に発行 I D を更新する（図 2 2 参照）。

【 0 1 8 5 】

そして、ホスト 3 0 0 からのデータのバッファメモリ 1 0 6 への格納の完了を待ち（ステップ S 4 0 5）、全てのデータがバッファメモリ 1 0 6 に格納されたことが通信部 3 2 2 から通知されると、処理シーケンサ 3 2 1 は、I O スケジューラ 1 0 4 に対して、処理の種類（この場合には書き込み）と、処理の識別 I D と、処理の対象となる記憶媒体 1 0 1 中の領域を示す情報と、処理の対象となるバッファメモリ 1 0 6 の領域を示す情報とを登録する。I O スケジューラ 1 0 4 は、書き込み要求に応じて、媒体制御部 1 0 5 に、書き込み対象のデータの書き込み指示を行う。媒体制御部 1 0 5 は、登録内容に応じてバッファメモリ 1 0 6 から記憶媒体 1 0 1 へのデータの書き込み処理を行う（ステップ S 4 0 6）。

【 0 1 8 6 】

そして、ストレージ 3 0 2 から準備完了のメッセージ（遅延 write コマンドに

対する応答）が送信されるのを待つ（ステップS407）。ストレージ302からの準備完了のメッセージを受信したことが通信部322から通知されると、処理シーケンサ321は、通信部322に、バッファメモリ106に格納されたデータをストレージ302に送信させる（ステップS408）。その後、ストレージ302から受信完了のメッセージが送信されるのと、媒体制御部105からの書き込み完了通知とを待ち（ステップS409）、ストレージ302からの受信完了のメッセージを受信したことが通信部322から通知され、かつ、媒体制御部105からの書き込み完了通知を受けると、ホスト300に完了通知し（ステップS410）、処理を終了する。

【0187】

ステップS408において送信するデータの順番は、対応する遅延write コマンドの発行順序と異なってもよい。ストレージ301がホスト300から連続してwrite コマンドを受信し、連続して遅延write コマンドを発行したとする。この場合、ストレージ301は、先に発行した遅延write コマンドに対応するデータの送信（ステップS408）が完了しないうちに、後の遅延write コマンドに対応するデータの送信（ステップS408）を開始してもよい。

【0188】

次に、ストレージ302が遅延write コマンドを受信したときの処理を説明する。図24は、ストレージ302のストレージコントローラ320における処理シーケンサ321が遅延write コマンドを受信したときの動作を示すフローチャートである。

【0189】

処理シーケンサ321は、まず、受け取るデータに必要な領域を同期用バッファメモリ323に確保し（ステップS420）、要求の送信元（この例ではストレージ301）に受信準備完了の通知を送るように通信部322に指示する。通信部322は、指示に応じて受信準備完了のメッセージを送信元にする（ステップS421）。次いで、データの到着を待ち（ステップS422）、データが到着し通信部322からデータの格納位置の問い合わせを受けると、ステップS420で確保した同期用バッファメモリ323の領域を通信部322に通知する（

ステップ S 4 2 3)。そして、要求の送信先からのデータの同期用バッファメモリ 3 2 3 への格納が完了するのを待ち（ステップ S 4 2 4）、データの格納が完了したら、要求の送信先に遅延 write コマンドにもとづく処理を完了したことを通知するように通信部 3 2 2 に指示する。通信部 3 2 2 は、指示に応じて、遅延 write コマンドにもとづく処理を完了したことを要求の送信先に通知する（ステップ S 4 2 5）。そして、処理シーケンサ 3 2 1 は処理を終了する。

【 0 1 9 0 】

また、処理シーケンサ 3 2 1 は、データを記憶媒体 1 0 1 に書き込む際に、同期用バッファメモリ 3 2 3 に既に格納されているデータと同一の同期 ID を有し、かつ、格納する領域が一部だけでも重なるデータについては、既に同期用バッファメモリ 3 2 3 に格納されている重なる部分を書き込まないように制御する。すなわち、同じ領域に書き込むデータに関して、後続の遅延 write コマンドにもとづく処理でのデータが有効になるように制御を行う。

【 0 1 9 1 】

例えば、記憶媒体 1 0 1 中のある領域へのデータの書き込みを指示する遅延 write コマンドがストレージ 3 0 2 に届き、同じ領域への書き込みを指示する遅延 write コマンドが届いたとする。そして、この二つの遅延 write コマンドの同期 ID が同一であるとする。この場合、正常系のストレージ 3 0 1 において再開可能ポイントの間に、一回データが書き込まれ、さらにデータが上書きされたことを意味する。従って、処理シーケンサ 3 2 1 は、同じ領域に書き込まれるデータのうち最初に書き込まれるデータを待機系ストレージ 3 0 2 の同期用バッファメモリ 3 2 3 の中に保持しなくてもよい。処理シーケンサ 3 2 1 は、新たにストレージ 3 0 1 から遅延 write コマンドを受信したときに、上書きされることになるデータが同期用バッファメモリ 3 2 3 に存在するか否かを判断する。そして、上書きされることになるデータを特定したならばそのデータを同期用バッファメモリ 3 2 3 から削除する。

【 0 1 9 2 】

次に、アプリケーション 3 1 0 a, 3 1 0 b が、ストレージ 3 0 1 に再開可能ポイント通知を通知する再開可能ポイント通知処理の動作を説明する。ここでは

、アプリケーション 3 1 0 a が再開可能ポイント通知処理を行う場合を例にする。

【 0 1 9 3 】

アプリケーション 3 1 0 a が、再開可能ポイントを通知する場合、I O 管理部 3 1 1 に再開可能ポイントを通知するように指示する。I O 管理部 3 1 1 において、アプリケーション 3 1 0 a から再開可能ポイントの通知が指示されると、再開可能ポイント通知部 3 1 2 が、ストレージ 3 0 1 に対して再開可能ポイント通知コマンドを発行する。

【 0 1 9 4 】

ストレージ 3 0 1 に再開可能ポイント通知コマンドが到着すると、ストレージ 3 0 1 のストレージコントローラ 3 2 0 における通信部 3 2 2 に再開可能ポイント通知コマンドが入力される。通信部 3 2 2 は、再開可能ポイント通知コマンドを受け取ると、処理シーケンサ 3 2 1 に再開可能ポイント通知コマンドを渡し、再開可能ポイント通知処理の開始を指示する。

【 0 1 9 5 】

図 2 5 は、処理シーケンサ 3 2 1 の動作を示すフローチャートである。処理シーケンサ 3 2 1 は、まず、その時点での内部の同期 I D の値を保持した後に、同期 I D の値を更新する（ステップ S 4 4 0）。同期 I D の値を保持するとは、例えば、レジスタに保存することである。次いで、ステップ S 4 4 0 で保持した同期 I D を指定した遅延データ反映コマンドをストレージ 3 0 2 に送信するように通信部 3 2 2 に指示する。通信部 3 2 2 は、指示に応じて、遅延データ反映コマンドをストレージ 3 0 2 に送信する（ステップ S 4 4 1）。そして、ストレージ 3 0 2 から遅延データ反映コマンドにもとづく処理の完了のメッセージが送られてくるのを待つ（ステップ S 4 4 2）。ストレージ 3 0 2 から完了のメッセージが到着したことが通信部 3 2 2 から通されたら、通信部 3 2 2 を介してホスト 3 0 0 に再開可能ポイント通知コマンドにもとづく処理の完了を通知し（ステップ S 4 4 3）、処理を終了する。

【 0 1 9 6 】

なお、図 2 2 に示すように、一つの遅延データ反映コマンドの送信後に出力す

る各遅延write コマンドには、レジスタ等に保存した同期 I D ではなく、更新後の同期 I D を付加する。

【 0 1 9 7 】

次に、ストレージ 3 0 2 の遅延データ反映コマンドを受けた際の動作を説明する。ストレージ 3 0 2 に遅延データ反映コマンドが到着すると、ストレージ 3 0 2 のストレージコントローラ 3 2 0 の通信部 3 2 2 に遅延データ反映コマンドが入力される。通信部 3 2 2 は、遅延データ反映コマンドを受け取ると、処理シーケンサ 3 2 1 に遅延データ反映コマンドを渡し、遅延データ反映処理の開始を指示する。

【 0 1 9 8 】

図 2 6 は、ストレージ 3 0 2 のストレージコントローラ 3 2 0 における処理シーケンサ 3 2 1 の遅延データ反映処理の動作を示すフローチャートである。処理シーケンサ 3 2 1 は、到着した遅延データ反映コマンドに付加された同期 I D の値が前回処理した遅延データ反映コマンドの次の値かどうかを判定する（ステップ S 4 6 0）。次の値でなかった場合には、同期 I D が、前回処理した遅延データ反映コマンドの同期 I D の次の値になっている遅延データ反映コマンドの到着を待つ。例えば、図 2 2 に示す例において、前回処理した遅延データ反映コマンドが、同期 I D 「 2 3 」の遅延データ反映コマンドであったとする。その後、同期 I D 「 2 5 」の遅延データ反映コマンドを受信した場合、同期 I D 「 2 4 」の遅延データ反映コマンドの到着を待つ。そして、同期 I D 「 2 4 」の遅延データ反映コマンドについてステップ S 4 6 1 ～ S 4 6 8 の処理を行った後、同期 I D 「 2 5 」の遅延データ反映コマンドについてステップ S 4 6 1 以降の処理を行う。

【 0 1 9 9 】

同期 I D が前回の同期 I D の次の値であった場合には、保持してある前回処理した遅延データ反映コマンドの発行 I D の値と、今回処理する遅延データ反映コマンドに付加された発行 I D の値の間の値の発行 I D を持つ遅延write コマンドに対応するデータが全て同期用バッファメモリ 3 2 3 に記録されているか否か検索する（ステップ S 4 6 1）。例えば、図 2 2 に示す各コマンドのうち、前回処

理した遅延データ反映コマンドの発行IDが「71」であり、今回処理する遅延データ反映コマンドに付加された発行IDが「76」であるとする。この場合、発行IDが「71」～「75」の遅延write コマンドに対応するデータが全て同期用バッファメモリ323に記録されているか否かを確認する。ただし、上書きされるデータであるとして削除したデータは記録確認の対象に含めなくてよい。抜けがあった場合には、各遅延write コマンドに対応するデータが全て待機系ストレージ302に到着し、同期用バッファメモリ323に記録されるまで待つ。

【0200】

ステップS463では、処理シーケンサ321は、同期用バッファメモリ323を検索し、遅延データ反映コマンドにより指定された同期IDと一致し、かつ、書き込み処理中でない遅延write コマンドを検索する。例えば、図22に示す発行ID「76」の遅延データ反映コマンドを受信して処理を行っている場合、そのコマンドに付加された同期ID「24」と一致し、かつ、書き込み処理中でない遅延write コマンドを検索する。検索した結果、見つからなければ、記憶媒体101へのデータの書き込み処理が開始されていない遅延write コマンドが存在しないことになる。この場合、ステップS465に移行する。また、検索対象の遅延write コマンドが見つかった場合には、記憶媒体101へのデータの書き込み処理が開始されていない遅延write コマンドが存在することになる。この場合、ステップS464に移行する。

【0201】

ステップS464では、ステップS462の検索で見つかった遅延write コマンドで指定された記憶媒体101上のデータを書き込むべき場所（オフセットアドレス）に対して、同期用バッファメモリ323中の対応したデータを書き込む指示（書き込み要求）をI/Oスケジューラ104に登録する。I/Oスケジューラ104は、書き込み要求に応じて、媒体制御部105に、データ転送対象のデータの書き込み指示を行う。媒体制御部105は、書き込み指示に従って、データを同期用バッファメモリ323から記憶媒体101に出力させる。そして、処理シーケンサ321は、同期用バッファメモリ323中の検索された領域を書き込み処理中とし、ステップS462に戻る。

【0202】

ステップS465では、同期用バッファメモリ323から記憶媒体101ヘデータ出力処理を行っているか否かを判断する。そして、処理中であればステップS466に移行し、すでに処理が完了しているならばステップS468に移行する。ステップS466では、記憶媒体101ヘデータ出力処理のうち1つが終了するまで待ち、完了した処理の対象であった同期用バッファメモリ323中の領域を未使用状態にして（ステップS467）、ステップS465に戻る。ステップS468では、遅延データ反映コマンドの送信元（この例ではストレージ301）に遅延データ反映コマンドにもとづく処理の完了を通知するように通信部322に指示する。通信部322は、指示に応じて、処理の完了を遅延データ反映コマンドの送信元に通知する。また、処理シーケンサ321は、前回処理した遅延データ反映コマンドの情報として保持している同期IDおよび発行IDを、処理した遅延データ反映コマンドのものに更新し、処理を終了する。

【0203】

次に、災害発生時の動作を説明する。災害等によってホスト300が使用できなくなった場合には、ストレージ302を待機系から正常系にする。また、ホスト303が、ホスト300から処理を引き継ぐ。

【0204】

障害検知からアプリケーション再開までのホスト303の動作を説明する。まず、ホスト303中のホスト監視部313がホスト300の異常を検出する。ホスト監視部313は、ホスト300の異常を検出すると、待機系であるストレージ302に対して遅延データ破棄コマンドを発行する。そして、遅延データ破棄コマンドに対する応答を待ち、応答を受けたら、待機系ホスト303のホスト監視部313は、待機系ホストのアプリケーション310a、310bを実行させる。以後、ホスト303からのデータの書き込みおよび読み出しは、ストレージ302に対して行われる。

【0205】

なお、ホスト303中のホスト監視部313は、ホスト300と常時あるいは定期的に通信を行っている。ホスト監視部313が、一定時間ホスト300と通

信できなくなった場合、あるいは、ホスト 3 0 0 から異常が報告された場合に、ホスト監視部 3 1 3 は、ホスト 3 0 0 において災害が発生したと認識する。ホスト 3 0 3 がホスト 3 0 0 の異常を確実に認識するために、ホスト 3 0 3 とホスト 3 0 0 とは専用回線で接続されていることが好ましい。

【 0 2 0 6 】

次に、待機系ストレージ 3 0 2 が、待機系ホスト 3 0 3 から遅延データ破棄コマンドを受けた場合の動作を説明する。ストレージ 3 0 2 に遅延データ破棄コマンドが到着すると、ストレージ 3 0 2 のストレージコントローラ 3 2 0 における通信部 3 2 2 に遅延データ破棄コマンドが入力される。通信部 3 2 2 は、遅延データ破棄コマンドを受け取ると、処理シーケンサ 3 2 2 に遅延データ破棄コマンドを渡し、遅延データ破棄処理の開始を指示する。

【 0 2 0 7 】

図 2 7 は、処理シーケンサ 3 2 2 が実行する遅延データ破棄処理を示すフローチャートである。処理シーケンサ 3 2 2 は、同期用バッファメモリ 3 2 3 を検索する（ステップ S 4 8 0）。ステップ S 4 8 0 では、未だ遅延データ反映コマンドが到着していないため記憶媒体 1 0 1 に移動させる必要がないデータを検索する。処理シーケンサ 3 2 2 は、検索した結果、見つからなかった場合ステップ S 4 8 3 に移行し、見つかった場合にはステップ S 4 8 2 に移行する（ステップ S 4 8 1）。例えば、図 2 2 に示す各コマンドのうち、発行 ID 「7 6」の遅延データ反映コマンドを受信しておらず、発行 ID 「7 2」～「7 5」のうちの一部の遅延 write コマンドのデータが同期用バッファメモリ 3 2 3 に記憶されているならば、ステップ S 4 8 2 に移行する。発行 ID 「7 6」の遅延データ反映コマンドを受信して、各遅延 write コマンドのデータを記憶媒体 1 0 1 に移行している場合には、ステップ S 4 8 3 に移行する。

【 0 2 0 8 】

ステップ S 4 8 2 では、処理シーケンサ 3 2 2 は、ステップ S 4 8 0 で見つかったデータを記録している同期用バッファメモリ 3 2 3 中の領域を未使用状態にしてステップ S 4 8 0 に戻る。すると、次のステップ S 4 8 1 では、ステップ S 4 8 3 に移行することになる。ステップ S 4 8 3 では、遅延データ反映処理中で

あった場合にはステップ S 4 8 4 に移行し、遅延データ反映処理中でなかった場合にはステップ S 4 8 5 に移行する。

【 0 2 0 9 】

ステップ S 4 8 4 では、処理中の遅延データ反映処理の完了を持ち、完了したらステップ S 4 8 5 に移行する。ステップ S 4 8 5 では、遅延データ破棄コマンドの発行元（この例ではホスト 3 0 3）に遅延データ破棄コマンドにもとづく処理の完了を通知するように通信部 3 2 2 に指示する。通信部 3 2 2 は、指示に応じて、処理の完了を遅延データ破棄コマンドの発行元に通知する、そして、処理シーケンサ 3 2 2 は処理を終了する。

【 0 2 1 0 】

本実施の形態では、正常系のホスト 3 0 0 がストレージ 3 0 1 に write コマンドを出力したとき、待機系のストレージ 3 0 2 はストレージ 3 0 1 から書き込まれるデータを受信するが、記憶媒体 1 0 1 には記録せず、同期用バッファメモリ 3 2 3 に記録する。そして、ストレージ 3 0 2 は、再開可能ポイント通知コマンドを受信したときに、そのデータを同期用バッファメモリ 3 2 3 から記憶媒体 1 0 1 に移動させ、遅延データ破棄コマンドを受信した場合、同期用バッファメモリ 3 2 3 を未使用状態にする。従って、ストレージ 3 0 2 の記憶媒体 1 0 1 は、常にアプリケーション 3 1 0 a, 3 1 0 b の処理を再開できる状態に保たれる。その結果、ホスト 3 0 0 に異常が生じたときに、即座にホスト 3 0 3 が処理を続行することができる。

【 0 2 1 1 】

また、正常系ストレージ 3 0 1 は、再開可能ポイントと再開可能ポイントとの間で記憶媒体 1 0 1 に書き込んだデータを一度に待機系ストレージ 3 0 2 に送信するのではなく、write コマンドを受信したタイミング毎に送信する。従って、大量のデータを一度に送信しないので、ホスト 3 0 2 とのデータ転送時間が少なくてすむ。

【 0 2 1 2 】

さらに、待機系ストレージ 3 0 2 がストレージ 3 0 1 から受信する各コマンドは同期 ID および発行 ID によって管理され、再開可能ポイントと再開可能ポイ

ントとの間に正常系ストレージが送信したデータの到着順序は任意の順序でよい。従って、遅延write コマンドに対応するデータを遅延write コマンドが送信する順序が制限されないので、設計を行いやすくなる。

【 0 2 1 3 】

本実施の形態では、遅延write コマンドおよび遅延データ反映コマンドに同期 I D および発行 I D の双方を付加する場合を説明したが、発行 I D のみを付加するようにしてもよい。この場合、待機系ストレージ 3 0 2 の処理シーケンサ 3 2 1 は、遅延データ反映コマンドを受信した場合、ステップ S 4 6 0、S 4 6 1 の代わりに以下の処理を行えばよい。処理シーケンサ 3 2 1 は、遅延データ反映コマンドを受信した場合、前回処理した遅延データ反映コマンドの発行 I D と受信した遅延データ反映コマンドの発行 I D との間の発行 I D が付加された各コマンドを全て受信しているか否かを確認する。そして、各発行 I D が付加されたコマンドを全て受信してれば、ステップ S 4 6 2（図 2 6 参照）以降の処理を行う。各発行 I D が付加されたコマンドを全て受信していなければ、各コマンドを全て受信したときにステップ S 4 6 2 以降の処理を開始する。

【 0 2 1 4 】

例えば、図 2 2 に示す各コマンド（同期 I D は付加されていないものとする）のうち、発行 I D 「7 1」の遅延データ反映コマンドが前回処理した遅延データ反映コマンドであったとする。その後、発行 I D 「7 6」の遅延データ反映コマンドを受信した場合、発行 I D 「7 2」～「7 5」の各コマンドを受信しているか否かを確認し、これらのコマンドを全て受信した後にステップ S 4 6 2 以降の処理を開始する。また、発行 I D 「7 6」より先に発行 I D 「8 0」の遅延データ反映コマンドを受信した場合、制御シーケンサ 3 2 1 は、発行 I D 「7 2」～「7 9」の各コマンドを受信しているか否かを確認する。そして、発行 I D 「7 2」～「7 6」の各コマンドを全て受信したときに、発行 I D 「7 6」の遅延データ反映コマンドの処理を行う。その後、発行 I D 「7 7」～「7 9」のコマンドを全て受信したときに発行 I D 「8 0」の遅延データ反映コマンドの処理を行う。

【 0 2 1 5 】

なお、各コマンドに同期 I D も付加される場合には、遅延 write コマンドを受信したときに、上書きされるデータを特定することができる。一方、同期 I D を用いない場合には、処理シーケンサ 3 2 1 はステップ S 4 6 2 の処理を開始する直前に、上書きされるデータを特定し、そのデータを同期用バッファメモリ 3 2 3 から削除する。

【 0 2 1 6 】

第 7 の実施の形態において、遅延書き込み要求手段および書き込み実行要求手段は、ストレージ 3 0 1 の処理シーケンサ 3 2 1 および通信部 3 2 2 によって実現される。一時記憶手段は、同期用バッファメモリ 3 2 3 によって実現される。格納処理手段は、ストレージ 3 0 2 の処理シーケンサ 3 2 1、I O スケジューラ 1 0 4 および媒体制御部 1 0 5 によって実現される。再開可能ポイント通知手段は、再開可能ポイント通知部 3 1 2 によって実現される。

【 0 2 1 7 】

実施の形態 8.

図 2 8 は、本発明によるデータ複製システムの第 8 の実施の形態を示すブロック図である。この実施の形態では、ストレージが主体となってデータ転送を実行するとともに、正常系システムにおけるストレージ内のデータが遠隔地にミラーリングされる。さらに、システムには、少なくとも 1 世代前のスナップショットを保存できる待機系のストレージが設けられる。1 世代前のスナップショットとは、直近の再開可能ポイントにおいてデータが格納されていた記憶媒体 1 0 1 のアドレスの情報である。

【 0 2 1 8 】

図 2 8 に示すデータ複製システムにおいて、ストレージ 4 0 0 が、ストレージ 4 0 0 を使用するホスト 3 0 0 とローカルに接続されている。また、ストレージ 4 0 1 が、ストレージ 4 0 1 を使用するホスト 3 0 3 とローカルに接続されている。ストレージ 3 0 1 は、ネットワーク 1 3 を介してストレージ 3 0 2 に接続されている。また、ホスト 3 0 0 とホスト 3 0 3 とは、ホスト 3 0 3 がホスト 3 0 0 の状態を監視するために通信可能に接続されている。ホスト 3 0 0 とホスト 3 0 3 は専用回線によって接続されていることが好ましいが、専用回線以外のネッ

トワーク（例えばインターネット等）によって接続されていてもよい。なお、ホスト 3 0 0, 3 0 3 の構成は、第 7 の実施の形態におけるホスト 3 0 0, 3 0 3 の構成と同じである（図 2 0 参照）。

【 0 2 1 9 】

ストレージ 4 0 0, 4 0 1 は、例えば、単体の磁気ディスク装置、光ディスク装置または光磁気ディスク装置である。ストレージ 4 0 0, 4 0 1 として、単体の磁気ディスク装置、光ディスク装置または光磁気ディスク装置の集合であるディスクアレイ装置を使用することもできる。ホスト 3 0 0, 3 0 3 とストレージ 4 0 0, 4 0 1 とは、S C S I、ファイバチャネル（Fibre channel）、イーサネット（登録商標）等で接続される。なお、図 2 8 に示すシステムにおいて、ホスト 3 0 0 が、システムに障害が発生していないときに稼働する正常系ホストであり、ホスト 3 0 3 が、ホスト 3 0 0 において障害が発生したときに稼働する待機系ホストであるとする。

【 0 2 2 0 】

この実施の形態では、ホスト 3 0 0 上で動作するアプリケーション 3 1 0 a, 3 1 0 b が、動作中に、そのデータの状態であればアプリケーションがそのまま動作を再開可能なポイントで、ストレージ 3 0 1 に再開可能なポイントであることを知らせるために再開可能ポイント通知処理を行う。なお、アプリケーション 3 1 0 a, 3 1 0 b が、ストレージ 4 0 0 からデータを読み出す際のホスト 3 0 0 およびストレージ 4 0 0 の動作は、通常の方法読み出し処理の場合の動作と同様である。また、ストレージ 4 0 0 において、I O スケジューラ 1 0 4 および媒体制御部 1 0 5 の動作は、第 1 の実施の形態における I O スケジューラ 1 0 4 および媒体制御部 1 0 5 の動作と同様である。

【 0 2 2 1 】

本実施の形態におけるアプリケーション 3 1 0 a, 3 1 0 b は、再開機能を有するアプリケーションである。すなわち、ストレージの記憶媒体 1 0 1 のデータ記録状態が所定の状態になっていれば処理を再開できるアプリケーションである。

【 0 2 2 2 】

図29は、図28に示すストレージ400の構成例を示すブロック図である。なお、ストレージ401の構成も、図29に示すような構成である。図29に示すように、ストレージ400は、ストレージコントローラ410とストレージ本体である記憶媒体101とを含む。ストレージコントローラ410は、ホスト300および他のストレージと通信を行う通信部412、各処理のシーケンスを管理する処理シーケンサ411、記憶媒体101に対する処理命令の順序制御を行うI/Oスケジューラ104、I/Oスケジューラ104が発行する処理命令に従って記憶媒体101の動作を制御する媒体処理部105、ホスト300から記憶媒体101へのデータおよび記憶媒体101からホスト300へのデータを一時記憶するバッファメモリ106、論理ブロックアドレスの管理を行うLBA管理部413、および論理ブロックアドレスの管理に用いられるアドレステーブルを記憶したアドレステーブル記憶部414を含む。処理シーケンサ321は、例えば、プログラムに従って動作するCPUで実現される。

【0223】

なお、同期用バッファメモリ322として半導体メモリが使用される場合もあるし、磁気ディスク装置、光ディスク装置または光磁気ディスク装置等のより大容量の記憶装置が使用される場合もある。

【0224】

ストレージ400、401は、記憶媒体101の物理アドレス空間によって、データを管理する。一方、他の装置からデータの書き込み処理や読み込み処理等の要求を受け付ける場合には、書き込み領域や読み込み領域を論理アドレス空間のアドレスによって指定される。再開可能ポイントにおける論理アドレス空間をスナップショット論理アドレス空間と呼ぶ。また、最新の時点において論理アドレスを最新論理アドレス空間と呼ぶ。ストレージ400、401は、スナップショット論理アドレス空間の情報と、最新論理アドレス空間の情報とを用いて処理を進める。また、ストレージ400、401は、複数のスナップショット論理アドレス空間を管理する場合もある。

【0225】

アドレステーブル記憶部414が記憶するアドレステーブルの一例を図30に

示す。アドレステーブル 4 2 0 は、複数のエントリ 4 2 0 - 0 ~ 4 2 0 - n から構成されている。各エントリ 4 2 0 - 0 ~ 4 2 0 - n は、論理アドレス空間を固定長のデータに分割した論理ブロックに対応する。

【 0 2 2 6 】

図 3 1 (a) は、エントリの説明図である。各エントリ 4 2 0 - 0 ~ 4 2 0 - n を構成する情報には、最新論理アドレス空間に対応したストレージ内の物理ブロック番号と、スナップショット論理アドレス空間に対応した物理ブロック番号がある。各物理ブロック番号は、直近にスナップショットを作成してからブロック番号の変更があったことを記録するフラグとセットになっている。なお、物理ブロック番号とは、ストレージ 4 0 0 , 4 0 1 のアドレス空間を固定長のデータに分割した際の物理ブロック個々につけられた一意の番号である。なお、スナップショット論理アドレス空間に対応する物理ブロック番号のフラグは設けなくてもよい。

【 0 2 2 7 】

図 3 1 (b) は、一つのエントリの初期状態の例を示す。図 3 1 (b) では、「 a a a 」という物理ブロック番号に対応するアドレスにデータが格納されていることを示す。また、再開可能ポイント以降、データが変更されていないので、最新論理アドレス空間に対応する物理ブロック番号も「 a a a 」となっている。

「 a a a 」という物理ブロック番号に対応するアドレスへの書き込みが要求されたとする。その場合、書き込みは他のアドレスに行い、そのアドレスに対応する物理ブロック番号「 b b b 」を最新論理アドレス空間に対応する物理ブロック番号として保持する。また、新論理アドレス空間に対応するフラグも「変更有り」に更新する。その後、再開可能ポイントになったならば、図 3 1 (d) に示すように最新論理アドレス空間に対応する情報を、スナップショット論理アドレス空間に対応する情報にコピーする。また、最新論理アドレス空間に対応するフラグを「変更無し」とする。

【 0 2 2 8 】

本実施の形態では、スナップショット論理アドレス空間に対応する情報をスナップショットとして用いる。

【 0 2 2 9 】

この結果、次に再開可能ポイントとなるまでは、新たに書き込もうとしたデータは、書き込もうとしたアドレスとは別のアドレスに書き込まれる。従って、再開可能ポイントの時点におけるデータが残ることになる。待機系のストレージへのデータの書き込みをこのように管理することにより、ホスト 3 0 0 に異常が生じて、すぐにホスト 3 0 3 が処理を再開することができる。

【 0 2 3 0 】

また、アドレステーブル記憶部 4 1 4 は、どの論理アドレス空間にも含まれない物理ブロックのブロック番号が記録された未使用ブロックテーブルも記憶する。なお、アドレステーブル記憶部 4 1 4 として、不揮発性の半導体メモリ、磁気ディスク装置、光ディスク装置または光磁気ディスク等が使用される。また、アドレステーブル記憶部 4 1 4 として、記憶媒体 1 0 1 の記憶領域の一部が使用される場合もある。

【 0 2 3 1 】

次に、ホスト 3 0 0 が、ストレージ 4 0 0、4 0 1 にデータを書き込むときの動作を説明する。ホスト 3 0 0 は、ストレージ 4 0 0 にデータを書き込むときに、ストレージ 4 0 0 のストレージコントローラ 4 1 0 における通信部 4 1 2 に対して write コマンドを出力する。通信部 4 1 2 は、write コマンドを受け取ると、write コマンドを処理シーケンサ 4 1 1 に渡し、write 処理を開始することを処理シーケンサ 4 1 1 に指示する。処理シーケンサ 4 1 1 が実行する書き込み処理は、データ転送先ストレージが設定されている場合としない場合とで異なる。なお、ストレージ 4 0 0 には、データ転送先としてストレージ 4 0 1 が設定され、ストレージ 4 0 1 には、データ転送先が設定されていない。

【 0 2 3 2 】

図 3 2 は、データ転送先が設定されている場合のストレージコントローラ 4 1 0 における処理シーケンサ 4 1 1 の動作を示すフローチャートである。すなわち、この実施の形態では、ストレージ 4 0 0 における処理シーケンサ 4 1 1 の動作を示すフローチャートである。データ転送先が設定されている場合には、処理シーケンサ 4 1 1 は、まず、受け取るデータに必要な領域を同期用バッファメモリ

3 2 3に確保する（ステップS 5 0 0）。また、write コマンドで指定された書き込み先の論理アドレスとwrite 処理であることを指定して、L B A管理部 4 1 3に論理アドレスを物理アドレスに変換させる（ステップS 5 0 1）。L B A管理部 4 1 3は、図 3 1に示す最新論理アドレス空間に対応する物理ブロック番号を割り当て、その物理ブロック番号に対応する物理アドレスを処理シーケンサ 4 1 1に返す。ステップS 5 0 1におけるL B A管理部 4 1 3の動作の詳細については後述する。

【 0 2 3 3 】

次いで、処理シーケンサ 4 1 1は、準備完了の通知をホスト 3 0 0に送るように通信部 4 1 2に指示する（ステップS 5 0 2）。通信部 4 1 2は、指示に応じて、ホスト 3 0 0に準備完了の通知を送信する。そして、ホスト 3 0 0からデータが届くのを待ち（ステップS 5 0 3）、データが届いて通信部 4 1 2からデータを格納すべきバッファメモリ 1 0 6の領域の問い合わせを受けると、ステップS 5 0 0で確保した領域を通信部 4 1 2に知らせる（ステップS 5 0 4）。また、再開可能ポイント通知処理を行っている場合にはステップS 5 1 3に移行し、行っていなかった場合にはステップS 5 0 6に移行する（ステップS 5 0 5）。ここで、「再開可能ポイント通知処理を行っている場合」とは、ホスト 3 0 0から再開可能ポイントを通知され、ストレージ 4 0 1に対して所定の処理を行っている場合を指す。具体的には、後述するステップS 5 6 2～S 5 6 4の処理を行っている場合にはステップS 5 1 3に移行し、行っていなかった場合にはステップS 5 0 6に移行する。

【 0 2 3 4 】

ステップS 5 0 6では、処理シーケンサ 4 1 1は、データ転送先のストレージ（この例ではストレージ 4 0 1）に対してwrite コマンドを発行するように通信部 4 1 2に指示する。このwrite コマンドでは、書き込み先を論理アドレスで指定する。通信部 4 1 2は、指示に応じて、write コマンドをデータ転送先のストレージに送信する。そして、ホスト 3 0 0からのデータのバッファメモリ 1 0 6への格納の完了を待ち（ステップS 5 0 7）、全てのデータがバッファメモリ 1 0 6に格納されたことが通信部 4 1 2から通知されると、処理シーケンサ 4 1 1

は、ＩＯスケジューラ１０４に対して、ステップＳ５０１で変換した物理アドレスに対応した領域に、バッファメモリ１０６に格納されたデータを書き込む指示（書き込み要求）を登録する。ＩＯスケジューラ１０４は、書き込み要求に応じて、媒体制御部１０５に、書き込み対象のデータの書き込み指示を行う。媒体制御部１０５は、登録内容に応じてバッファメモリ１０６から記憶媒体１０１へのデータの書き込み処理を行う（ステップＳ５０８）。

【０２３５】

そして、ストレージ４０１から準備完了のメッセージが送信されるのを待ち（ステップＳ５０９）、ストレージ４０１からの準備完了のメッセージを受信したことが通信部４１２から通知されると、通信部４１２に、バッファメモリ１０６に格納されたデータをストレージ４０１に送信させる（ステップＳ５１０）。その後、ストレージ４０１から受信完了のメッセージが送信されるのと、媒体制御部１０５からの書き込み完了通知とを待ち（ステップＳ５１１）、ストレージ４０１からの受信完了のメッセージを受信したことが通信部４１２から通知され、かつ、媒体制御部１０５からの書き込み完了通知を受けると、ホスト３００に完了通知し（ステップＳ５１２）、処理を終了する。

【０２３６】

ステップＳ５１３では、ホスト３００からのデータのバッファメモリ１０６への格納の完了を待ち、全てのデータがバッファメモリ１０６に格納されたことが通信部４１２から通知されると、処理シーケンサ４１１は、ＩＯスケジューラ１０４に対して、ステップＳ５０１で変換した物理アドレスに対応した領域に、バッファメモリ１０６に格納されたデータを書き込む指示（書き込み要求）を登録する（ステップＳ５１４）。次いで、再開可能ポイント通知処理が完了するまで待ち（ステップＳ５１５）、再開可能ポイント通知処理が完了したら、データ転送先のストレージに対してwrite コマンドを発行するように通信部４１２に指示する（ステップＳ５１６）。このwrite コマンドでは、書き込み先を論理アドレスで指定する。ステップＳ５１６の後、ステップＳ５０９に移行する。

【０２３７】

なお、ステップＳ５０６～Ｓ５０８での処理と、ステップＳ５１３～Ｓ５１６

での処理では、ストレージ401に対するwrite コマンド発行処理と記憶媒体101への書き込み処理の順番が逆になっている。これは、以下の理由によるものである。ステップS506～S508では、ストレージ401にwrite コマンドを送信し、その応答が戻って来るまでの期間に記憶媒体101への書き込み処理を進めるため、write コマンド発行処理を記憶媒体101への書き込み処理よりも先に行うこととした。一方、ステップS513～S516では、再開可能ポイント通知処理の終了を待っている間、ストレージ401にエントリの変更を行わせている。従って、再開可能ポイント通知処理の終了まではストレージ401にwrite コマンドを送信できない。そこで、再開可能ポイント通知処理の終了を待つまでの間に記憶媒体101への書き込み処理を進めることとした。

【0238】

図33は、データ転送先が設定されている場合のストレージコントローラ410における処理シーケンサ411の動作を示すフローチャートである。すなわち、この実施の形態では、ストレージ401における処理シーケンサ411の動作を示すフローチャートである。より具体的には、ステップS506，S516で送信されたwrite コマンドを受信したストレージ401の処理シーケンサ411の動作を示すフローチャートである。

【0239】

処理シーケンサ411はwrite コマンドを受信すると、まず、受け取るデータに必要な領域をバッファメモリ106に確保する（ステップS520）。また、write コマンドで指定された書き込み先の論理アドレスとwrite 処理であることを指定して、LBA管理部413に論理アドレスを物理アドレスに変換させる（ステップS521）。この処理はステップS501（図32参照）と同様の処理である。

【0240】

次いで、処理シーケンサ411は、準備完了の通知をデータ転送元のストレージ（この例ではストレージ400）に送るように通信部412に指示する（ステップS522）。通信部412は、指示に応じて、データ転送元のストレージに準備完了の通知を送信する。そして、データ転送元のストレージからデータが届

くの待ち（ステップS523）、データが届いて通信部412からデータを格納すべきバッファメモリ106の領域の問い合わせを受けると、ステップS520で確保した領域を通信部412に知らせる（ステップS524）。

【0241】

さらに、データ転送元のストレージからのデータのバッファメモリ106への格納の完了を待ち（ステップS525）、全てのデータがバッファメモリ106に格納されたことが通信部412から通知されると、処理シーケンサ411は、I/Oスケジューラ104に対して、ステップS521で変換した物理アドレスに対応した領域に、バッファメモリ106に格納されたデータを書き込む指示（書き込み要求）を登録する。I/Oスケジューラ104は、書き込み要求に応じて、媒体制御部105に、書き込み対象のデータの書き込み指示を行う。媒体制御部105は、登録内容に応じてバッファメモリ106から記憶媒体101へのデータの書き込み処理を行う（ステップS526）。

【0242】

そして、媒体制御部105からの書き込み完了通知を待ち（ステップS527）、媒体制御部105からの書き込み完了通知を受けると、データ転送元のストレージに完了通知を行い（ステップS528）、処理を終了する。

【0243】

ストレージ400、401は、いずれも論理アドレスを指定されたwrite コマンドを受信して、書き込み処理を開始する。しかし、その論理アドレスに対応する物理アドレスにデータを書き込むのではなく、別の物理アドレス（ステップS501、S521で割り当てた物理アドレス）にデータを書き込む。従って、再開可能ポイントにおけるデータを記憶媒体101に保持し続けることができるので、正常系のホスト300に異常が生じて、待機系のホスト303は、短時間で処理を再開することができる。

【0244】

次に、LBA管理部413による論理アドレスから物理アドレスへの変換処理を説明する。なお、LBA管理部413による変換処理は、write 処理であることが指定された場合とされなかった場合とで異なる。write 処理であることが指

定されなかった場合には、LBA管理部413は、論理アドレスが指定されて変換処理の指示を受けると、論理アドレスから論理ブロック番号を算出し、アドレステーブル記憶部414から、最新論理アドレス空間でのその論理ブロック番号に対応した物理ブロック番号を取得する。また、物理ブロック番号と論理ブロックアドレスとから物理アドレスを算出し、算出した物理アドレスを処理シーケンサ411に通知する。なお、論理ブロック番号は、論理ブロックアドレスをブロック長で割った値（小数点以下は切り捨てる）である。また、物理アドレスを算出する場合、まず物理ブロックとブロック長の積（Pとする）を求める。また、論理アドレスをブロック長で割った値の余り（Qとする）を求める。PとQの和を物理アドレスとして算出する。

【0245】

図34は、write 処理であることが指定された場合のLBA管理部413の動作を示すフローチャートである。すなわち、ステップS501，S521におけるLBA管理部413の動作を示すフローチャートである。write 処理であることが指定された場合には、LBA管理部413は、論理アドレスから論理ブロック番号を算出する（ステップS540）、次いで、アドレステーブル記憶部414から、ステップS540で算出した論理ブロック番号に対応したエントリを取得する（ステップS541）。ステップS541では、最新論理アドレス空間におけるエントリを取得する。そして、エントリ中の、最新論理アドレス空間に対応する物理ブロック番号と対になるフラグを確認する（ステップS542）。フラグが「変更有り」となっている場合にはステップS546に移行し、「変更無し」となっている場合にはステップS543に移行する。

【0246】

フラグが「変更無し」となっている場合とは、図31（b）に示すように、最新論理アドレス空間に対応する物理ブロック番号がスナップショット論理アドレス空間に対応する物理ブロック番号と等しい場合である。以下、図31（b），（c）を用いて説明する。ステップS543では、LBA管理部413は、アドレステーブル記憶部414中の未使用ブロックテーブルから未使用の物理ブロック番号を入手する（ステップS543）。ここでは、「bbb」という物理ブ

ック番号を入手したものとする。また、アドレステーブル記憶部 4 1 4 に登録されている未使用ブロックテーブルの情報から、入手した物理ブロック番号の情報を削除する。そして、削除後の未使用ブロックテーブルの情報をアドレステーブル記憶部 4 1 4 に登録し直す。

【 0 2 4 7 】

さらに、ステップ S 5 4 3 で入手した物理ブロック番号 (b b b) に該当する領域に、ステップ S 5 4 1 で入手したエントリ中の最新論理アドレス空間に対応する物理ブロック (a a a) に該当する領域のデータをコピーする旨の情報 (コピー要求) を I O スケジューラ 1 0 4 に登録する。I O スケジューラ 1 0 4 は、コピー要求に応じて、媒体制御部 1 0 5 に、対象のデータのコピー指示を行う。媒体制御部 1 0 5 は、記憶媒体 1 0 1 における指定された領域間のコピー処理を行う (ステップ S 5 4 4) 。コピー処理が完了したら、L B A 管理部 4 1 3 は、ステップ S 5 4 1 で入手したエントリの最新論理アドレス空間に対応する物理ブロック番号 (a a a) をステップ S 5 4 3 で入手した物理ブロック番号 (b b b) に変更し、変更後の物理ブロック番号とセットとなっているフラグを「変更有り」に更新する。また、アドレステーブル記憶部 4 1 4 に、変更したエントリを記憶させる (ステップ S 5 4 5) 。そして、ステップ S 5 4 6 に移行する。

【 0 2 4 8 】

ステップ S 5 4 6 では、L B A 管理部 4 1 3 は、エントリの最新論理アドレス空間に対応する物理ブロック番号から物理アドレスを算出し、算出した物理アドレスを処理シーケンサ 4 1 1 に通知し (ステップ S 5 4 7) 、処理を終了する。

【 0 2 4 9 】

ステップ S 5 4 2 の後、ステップ S 5 4 3 以降の処理を行うと、図 3 1 (b) に示す物理ブロック番号は、図 (c) に示すように変更される。しかし、ステップ S 5 4 7 までの処理が終了した時点では、「 b b b 」に対応する物理アドレスには、「 a a a 」に対応する物理アドレスと同一のデータが格納されている。正常系ストレージ 4 0 0 が「 b b b 」に対応する物理アドレスのデータを書き換えるのは、ステップ S 5 0 8 またはステップ S 5 1 4 においてである (図 3 2 参照) 。また、待機系ストレージ 4 0 1 が「 b b b 」に対応する物理アドレスのデー

タを書き換えるのはステップ S 5 2 6 においてである（図 3 3 参照）。

【 0 2 5 0 】

次に、アプリケーション 3 1 0 a, 3 1 0 b が、ストレージ 4 0 0 に再開可能ポイント通知を通知するときの動作を説明する。アプリケーション 3 1 0 a, 3 1 0 b は、ストレージ 4 0 0 に再開可能ポイント通知を通知するとき、I/O 管理部 3 1 1（図 2 0 参照）に再開可能ポイントの通知を指示する。すると、I/O 管理部 3 1 1 において再開可能ポイント通知部 3 1 2 は、ストレージ 4 0 0 に対して再開可能ポイント通知コマンドを発行する。

【 0 2 5 1 】

ストレージ 4 0 0 に再開可能ポイント通知コマンドが到着すると、再開可能ポイント通知コマンドは、ストレージコントローラ 4 1 0 における通信部 4 1 2 に入力される。通信部 4 1 2 は、再開可能ポイント通知コマンドを受け取ると、再開可能ポイント通知コマンドを処理シーケンサ 4 1 1 に渡し、再開可能ポイント通知処理を開始することを処理シーケンサ 4 1 1 に指示する。

【 0 2 5 2 】

図 3 5 は、再開可能ポイント通知コマンドを受け取った処理シーケンサ 4 1 1 の動作を示すフローチャートである。処理シーケンサ 4 1 1 は、再開可能ポイント通知コマンドを受け取ると、図 3 5 に示すように、まず、他ストレージに対して要求した write 処理のうちで完了していないものがあるか否かを調べ（ステップ S 5 6 0）、あった場合にはステップ S 5 6 1 に処理を実行する。すなわち、他ストレージに対して要求した write 処理が全て完了するのを待つ（ステップ S 5 6 1）。なお、処理シーケンサ 4 1 1 は、発行した各 write コマンドに対する応答が完了しているのか否かを示す一覧情報を管理する。ステップ S 5 6 0 では、この情報に基づいて判断を行えばよい。

【 0 2 5 3 】

処理シーケンサ 4 1 1 は、他ストレージに対して要求した write 処理が全て完了している状態で、通信部 4 1 2 に、データ転送先のストレージ（この例ではストレージ 4 0 1）にスナップショット作成コマンド（スナップショット作成要求）を送信するように指示する。通信部 4 1 2 は、指示に応じて、データ転送先の

ストレージにスナップショット作成コマンドを送信する（ステップ S 5 6 2）。スナップショット作成コマンドとは、最新論理アドレス空間に対応する情報を、スナップショット論理アドレス空間に対応する情報にコピーするようにし、かつ、最新論理アドレス空間に対応するフラグを「変更無し」に更新するように要求するコマンドである。処理シーケンサ 4 1 1 は、ステップ S 5 6 2 の後、転送先ストレージからの応答を待ち（ステップ S 5 6 3）、応答が到着ことが通信部 4 1 2 から通知されたら、通信部 4 1 2 にホスト 3 0 0 に対して完了を通知するように指示し（ステップ S 5 6 4）、処理を終了する。

【 0 2 5 4 】

次に、データ転送先のストレージであるストレージ 4 0 1 がスナップショット作成コマンドを受信した場合の動作を説明する。ストレージ 4 0 1 にスナップショット作成コマンドが到着すると、ストレージ 4 0 1 のストレージコントローラ 4 1 0 の通信部 4 1 2 にスナップショット作成コマンドが入力される。通信部 4 1 2 は、スナップショット作成コマンドを受け取ると、処理シーケンサ 4 1 1 にスナップショット作成コマンドを渡し、スナップショット作成処理の開始を指示する。

【 0 2 5 5 】

スナップショット作成処理では、処理シーケンサ 4 1 1 は、アドレステーブル記憶部 4 1 4 中のアドレステーブルの全エントリに対して以下の処理を行う。すなわち、最新論理アドレス空間に対応するフラグに「変更有り」が記録されていた場合には、そのエントリのスナップショット論理アドレス空間に対応する物理ブロック番号を、未使用ブロックとしてアドレステーブル記憶部 4 1 4 中の未使用ブロックテーブルに登録する。また、スナップショット論理アドレス空間に対応するブロック番号に最新論理アドレス空間に対応する物理ブロック番号をコピーする。このとき、フラグの値もコピーする。その後、最新論理アドレス空間に対応するフラグを「変更無し」に初期化する。

【 0 2 5 6 】

例えば、図 3 1（c）に示すエントリでは、最新論理アドレス空間に対応するフラグに「変更有り」が記録されている。この場合、スナップショット論理アド

レス空間に対応する物理ブロック番号「a a a」を未使用ブロックとして未使用ブロックテーブルに登録する。そして、スナップショット論理アドレス空間に対応するブロック番号に、最新論理アドレス空間に対応する物理ブロック番号「b b b」をコピーする。そして、最新論理アドレス空間に対応するフラグ「変更有り」も同様にコピーする。さらに、最新論理アドレス空間に対応するフラグを「変更無し」に初期化する。すると、エントリの状態は、図 3 1 (c) に示す状態から、図 3 1 (d) に示す状態になる。

【 0 2 5 7 】

また、最新論理アドレス空間に対応するフラグが「変更無し」である場合、直前の再開可能ポイント以降その物理ブロック番号に対応するアドレスに格納されたデータが変更されていないことを意味する。従って、そのエントリには何も処理を行わない。

【 0 2 5 8 】

最新論理アドレス空間に対応するフラグが「変更有り」となっている全エントリに対して処理を終了した後、処理シーケンサ 4 1 1 は、通信部 4 1 2 を用いてスナップショット作成コマンドに対する応答を送信する。

【 0 2 5 9 】

次に、災害発生時の動作を説明する。災害等によってホスト 3 0 0 が使用できなくなった場合には、ストレージ 4 0 1 を待機系から正常系にする。また、ホスト 3 0 3 が、ホスト 3 0 0 から処理を引き継ぐ。

【 0 2 6 0 】

障害検知からアプリケーション再開までのホスト 3 0 3 の動作を説明する。まず、ホスト 3 0 3 中のホスト監視部 3 1 3 (図 2 0 参照) がホスト 3 0 0 の異常を検出する。ホスト監視部 3 1 3 は、ホスト 3 0 0 の異常を検出すると、待機系であるストレージ 4 0 1 に対してスナップショット復帰コマンドを発行する。そして、スナップショット復帰コマンドに対する応答を待ち、応答を受けたら、ホスト監視部 3 1 3 は、アプリケーション 3 1 0 a, 3 1 0 b を実行させる。以後、ホスト 3 0 3 からのデータの書き込みおよび読み出しは、ストレージ 4 0 1 に対して行われる。

【 0 2 6 1 】

なお、ホスト 3 0 3 中のホスト監視部 3 1 3 は、ホスト 3 0 0 と常時あるいは定期的に通信を行っている。ホスト監視部 3 1 3 が、一定時間ホスト 3 0 0 と通信できなくなった場合、あるいは、ホスト 3 0 0 から異常が報告された場合に、ホスト監視部 3 1 3 は、ホスト 3 0 0 において災害が発生したと認識する。

【 0 2 6 2 】

次に、ストレージ 4 0 1 が、スナップショット復帰コマンドを受けた場合の動作を説明する。ストレージ 4 0 1 にスナップショット復帰コマンドが到着すると、ストレージ 4 0 1 のストレージコントローラ 4 1 0 における通信部 4 1 2 にスナップショット復帰コマンドが入力される。通信部 4 1 2 は、スナップショット復帰コマンドを受け取ると、処理シーケンサ 4 1 1 にスナップショット復帰コマンドを渡し、スナップショット復帰処理の開始を指示する。

【 0 2 6 3 】

スナップショット復帰処理では、処理シーケンサ 4 1 1 は、アドレステーブル記憶部中 4 1 4 のアドレステーブルの全エントリに対して以下の処理を行う。すなわち、最新論理アドレス空間に対応するフラグに「変更有り」が記録されていた場合には、その最新論理アドレス空間に対応する物理ブロック番号を未使用ブロックとしてアドレステーブル記憶部 4 1 4 の未使用ブロックテーブルに登録する。その結果、その物理ブロック番号に対応する領域は未使用状態として解放される。次いで、記憶媒体へのデータ格納状況を示す格納情報（最新論理アドレス空間に対応する物理ブロック番号およびフラグ）を、直前のスナップショット作成時の状態に戻す。すなわち、最新論理アドレス空間に対応するブロック番号にスナップショット論理アドレス空間に対応する物理ブロック番号をコピーする。また、最新論理アドレス空間に対応するフラグを「変更有り」から「変更無し」にする。

【 0 2 6 4 】

例えば、スナップショット復帰処理開始時に、あるエントリが図 3 1 (c) に示す状態であったとする。この最新論理アドレス空間に対応するフラグは「変更有り」となっている。従って、処理シーケンサ 4 1 1 は、その最新論理アドレス

空間に対応する物理ブロック番号「b b b」を未使用ブロックとして未使用ブロックテーブルに登録する。次いで、最新論理アドレス空間に対応するブロック番号「b b b」にスナップショット論理アドレス空間に対応する物理ブロック番号をコピーする。この結果、最新論理アドレス空間に対応するブロック番号は「a a a」になる。また、最新論理アドレス空間に対応するフラグを「変更有り」から「変更無し」にする。この結果、格納情報（最新論理アドレス空間に対応する物理ブロック番号およびフラグ）は、直前のスナップショット作成時の状態に戻る。このように、スナップショット復帰処理を行うことにより、エントリは直前の再開可能ポイントにおける状態に復帰する。

【 0 2 6 5 】

また、最新論理アドレス空間に対応するフラグが「変更無し」である場合、直前の再開可能ポイント以降その物理ブロック番号に対応するアドレスに格納されたデータが変更されていないことを意味する。従って、そのエントリには何も処理を行わない。

【 0 2 6 6 】

本実施の形態では、待機系のストレージ 4 0 1 はストレージ 4 0 0 から受信したデータを、新たに割り当てた物理ブロック番号に対応するアドレスに格納する。そして、ストレージ 4 0 1 は、再開可能ポイントにおいてスナップショット作成コマンドを受信すると、その新たに割り当てた物理ブロックの情報をスナップショットとして保持する。また、再開可能ポイントの前に、正常系のホストに異常が発生した場合には、新たに割り当てた物理ブロック番号を未使用状態にする。新たに割り当てた物理ブロック番号を未使用の状態にすれば、ホスト 3 0 3 は、アプリケーション 3 1 0 a, 3 1 0 b の処理を再開できる。すなわち、ホスト 3 0 0 に異常が生じたとしても、ホスト 3 0 3 が処理を再開するまでの時間は短くて済む。

【 0 2 6 7 】

また、正常系ストレージ 4 0 0 は、再開可能ポイントと再開可能ポイントとの間で記憶媒体 1 0 1 に書き込んだデータを一度に待機系ストレージ 4 0 1 に送信するのではなく、ホスト 3 0 0 から write コマンドを受信したタイミング毎に送

信する。従って、大量のデータを一度に送信しないので、ホスト 4 0 1 とのデータ転送時間が少なくてすむ。

【 0 2 6 8 】

次に、第 8 の実施の形態の変形例について説明する。図 3 2 および図 3 5 に示された処理例では、再開可能ポイント通知処理が行われていた場合には write コマンドの発行は待たされていた（ステップ S 5 0 5 参照）。また、write 処理のうちで完了していないものがある場合にはスナップショット作成コマンドの発行が待たされていた（ステップ S 5 6 0 参照）。これに対し、本変形例では、ストレージ 4 0 0 は、ホスト 3 0 0 から write コマンドを受信した場合、再開可能ポイント通知処理の状態によらずにストレージ 4 0 1 に write コマンドを送信する。すなわち、図 3 2 に示すステップ S 5 0 4 の後、即座にステップ S 5 0 6 以降の処理を開始する。また、ストレージ 4 0 0 は、ホスト 3 0 0 から再開可能ポイント通知を受けた場合には、write 処理の状況によらず即座にストレージ 4 0 1 にスナップショット作成コマンドが発行される。すなわち、再開可能ポイント通知を受けた場合、即座に図 3 5 に示すステップ S 5 6 2 以降の処理を開始する。

【 0 2 6 9 】

本変形例では、ストレージ 4 0 0 は、ストレージ 4 0 1 に対して送信する全てのコマンドに発行 ID を付加する。また、待機系のストレージ 4 0 1 は、処理済発行 ID 情報を保持する。処理済発行 ID 情報は、ストレージ 4 0 0 から受信した各コマンドのどのコマンドまでの処理が完了したのかを示す情報である。例えば、処理済発行 ID 情報の内容が「5 3」であるならば、発行 ID「5 3」までのコマンドに対する処理が完了したことを意味する。

【 0 2 7 0 】

図 3 6 は、第 8 の実施の形態の変形例において、write コマンドを受信したストレージ 4 0 1 の処理シーケンサ 4 1 1 の動作を示すフローチャートである。処理シーケンサ 4 1 1 は、まず、受け取るデータに必要な領域をバッファメモリ 1 0 6 に確保する（ステップ S 5 8 0）。また、write コマンドで指定された書き込み先の論理アドレスと write 処理であることを指定して、L B A 管理部 4 1 3 に論理アドレスを物理アドレスに変換させる（ステップ S 5 8 1）。ステップ S

5 8 1において、L B A管理部 4 1 3は、ステップ S 5 2 1の場合と同様にエントリに対して処理を行い、物理アドレスを処理シーケンサ 4 1 1に渡す。

【 0 2 7 1 】

次いで、処理シーケンサ 4 1 1は、準備完了の通知をデータ転送元のストレージ（この例ではストレージ 4 0 0）に送るように通信部 4 1 2に指示する（ステップ S 5 8 2）。通信部 4 1 2は、指示に応じて、データ転送元のストレージに準備完了の通知を送信する。そして、データ転送元のストレージからデータが届くのを待ち（ステップ S 5 8 3）、データが届いて通信部 4 1 2からデータを格納すべきバッファメモリ 1 0 6の領域の問い合わせを受けると、ステップ S 5 8 0で確保した領域を通信部 4 1 2に知らせる（ステップ S 5 8 4）。

【 0 2 7 2 】

さらに、データ転送元のストレージからのデータのバッファメモリ 1 0 6への格納の完了を待ち（ステップ S 5 8 5）、全てのデータがバッファメモリ 1 0 6に格納されたことが通信部 4 1 2から通知されると、処理シーケンサ 4 1 1は、処理済発行 I D情報の値が、処理対象の write コマンドに付加された発行 I Dの一つ前の値になるまで待つ（ステップ S 5 8 6）。例えば、発行 I D「5 4」の write コマンドを受信して、ステップ S 5 8 0からステップ S 5 8 5までの動作を行ったとする。このとき、発行 I D「5 3」までの各コマンド（write コマンドやスナップショット作成コマンド等）を受信しているとは限らない。この場合、ステップ S 5 8 6では、発行 I D「5 3」までの各コマンドを受信し、その各コマンドに応じた処理を全て完了させるまで、発行 I D「5 4」に対する処理を中断する。処理済発行 I D情報が「5 3」になったならば、発行 I D「5 4」に対する処理を再開し、ステップ S 5 8 7に移行する。

【 0 2 7 3 】

処理済発行 I D情報の値が発行 I Dの一つ前の値になったら、処理シーケンサ 4 1 1は、I Oスケジューラ 1 0 4に対して、ステップ S 5 8 1で変換した物理アドレスに対応した領域に、バッファメモリ 1 0 6に格納されたデータを書き込む指示（書き込み要求）を登録する。I Oスケジューラ 1 0 4は、書き込み要求に応じて、媒体制御部 1 0 5に、書き込み対象のデータの書き込み指示を行う。

媒体制御部 1 0 5 は、登録内容に応じてバッファメモリ 1 0 6 から記憶媒体 1 0 1 へのデータの書き込み処理を行う（ステップ S 5 8 7）。

【 0 2 7 4 】

そして、媒体制御部 1 0 5 からの書き込み完了通知を待ち（ステップ S 5 8 8）、媒体制御部 1 0 5 からの書き込み完了通知を受けると、データ転送元のストレージに完了通知を行う（ステップ S 5 8 9）。また、ステップ S 5 8 9 では、write コマンドに付加された発行 ID を処理済発行 ID 情報に反映させる。例えば、発行 ID 「5 4」の write コマンドについてステップ S 5 8 9 までの処理を行ったならば、処理済発行 ID 情報の情報を「5 3」から「5 4」に更新する。

【 0 2 7 5 】

また、第 8 の実施の形態の変形例では、ストレージ 4 0 1 がスナップショット作成コマンドを受信した場合には、以下のように動作する。すなわち、ストレージ 4 0 1 にスナップショット作成コマンドが到着すると、ストレージ 4 0 1 のストレージコントローラ 4 1 0 の通信部 4 1 2 にスナップショット作成コマンドが入力される。通信部 4 1 2 は、スナップショット作成コマンドを受け取ると、処理シーケンサ 4 1 1 にスナップショット作成コマンドを渡し、スナップショット作成処理の開始を指示する。

【 0 2 7 6 】

スナップショット作成処理では、処理済発行 ID 情報後が、処理対象の要求に付加された発行 ID の一つ前の値になるまで待つ。すなわち、ステップ S 5 8 6 の場合と同様に、スナップショット作成コマンドよりも一つ前に発行されたコマンドに応じた処理を全て完了させるまで、そのスナップショット作成コマンドに応じた処理を進めずに待つ。

【 0 2 7 7 】

処理済発行 ID 情報の値が発行 ID の一つ前の値になったら、処理シーケンサ 4 1 1 は、アドレステーブル記憶部中 4 1 4 のアドレステーブルの全エントリに対して以下の処理を行う。すなわち、最新論理アドレス空間に対応するフラグに「変更有り」が記録されていた場合には、その最新論理アドレス空間に対応する物理ブロック番号を未使用ブロックとしてアドレステーブル記憶部 4 1 4 の未使

用ブロックテーブルに登録する。次いで、最新論理アドレス空間に対応するブロック番号にスナップショット論理アドレス空間に対応する物理ブロック番号をコピーする。また、最新論理アドレス空間に対応するフラグを「変更有り」から「変更無し」にする。

【 0 2 7 8 】

また、最新論理アドレス空間に対応するフラグが「変更無し」である場合、直前の再開可能ポイント以降その物理ブロック番号に対応するアドレスに格納されたデータが変更されていないことを意味する。従って、そのエントリには何も処理を行わない。

【 0 2 7 9 】

以上の処理全てが終了したならば、処理シーケンサ 4 1 1 は、処理済発行 ID 情報の値をスナップショット作成コマンドに付加された発行 ID に変更する。そして、通信部 4 1 2 を用いてスナップショット作成コマンドに対する応答をストレージ 4 0 0 に送信する。

【 0 2 8 0 】

本変形例では、ストレージ 4 0 1 が発行 ID の順番に各コマンドの処理を進めていく点が、既に説明した第 8 の実施の形態と異なる。しかし、エントリに対する処理自体に相違点はない。従って、ホスト 3 0 0 に異常が生じたとしても、ホスト 3 0 3 が処理を再開するまでの時間は短くて済む。また、ストレージ 4 0 0 , 4 0 1 間でのデータ転送時間も短くて済む。

【 0 2 8 1 】

第 8 の実施の形態において、書き込み要求手段およびスナップショット作成要求手段は、ストレージ 4 0 0 の処理シーケンサ 4 1 1 および通信部 4 1 2 によって実現される。スナップショット作成手段は、ストレージ 4 0 1 の処理シーケンサ 4 1 1、LBA 管理部 4 1 3 およびアドレステーブル記憶部 4 1 4 によって実現される。再開可能ポイント通知手段は、ホスト 3 0 0 の再開可能ポイント通知部 3 1 2 によって実現される。

【 0 2 8 2 】

実施の形態 9.

図 3 7 は、本発明によるデータ複製システムの第 9 の実施の形態を示すブロック図である。図 3 7 に示すデータ複製システムにおいて、正常系のストレージ 4 0 0 が、ストレージ 4 0 0 を使用する正常系のホスト 6 0 0 とローカルに接続されている。また、待機系のストレージ 4 0 1 が、ストレージ 4 0 1 を使用する待機系のホスト 6 0 1 とローカルに接続されている。ストレージ 4 0 0 は、ネットワーク 1 3 を介してストレージ 4 0 1 に接続されている。また、ホスト 6 0 0 は、ネットワーク 6 0 2 を介してホスト 6 0 1 に接続されている。

【 0 2 8 3 】

なお、ストレージ 4 0 0、4 0 1 の構成および動作は、第 8 の実施の形態におけるスナップショット機能を有するストレージ 4 0 0、4 0 1 の構成および動作と同じである（図 2 9 参照）。また、ストレージ 4 0 0、4 0 1 に代えて、第 7 の実施の形態におけるストレージ 3 0 1、3 0 2（図 2 1 参照）を用いてもよい。

【 0 2 8 4 】

また、ネットワーク 1 3、6 0 2 は、1 つのネットワークであってもよい。ただし、ホスト 6 0 0、6 0 1 が接続されるネットワーク 6 0 2 は、専用回線とすることが好ましい。

【 0 2 8 5 】

図 3 8 は、ホスト 6 0 0 の構成例を示すブロック図である。図 3 8 に示すように、ホスト 6 0 0 において、単数あるいは複数のアプリケーションが動作する。ここでは、2 つのアプリケーション 6 0 3 a、6 0 3 b を例示する。アプリケーション 6 0 3 a、6 0 3 b は、I O 管理部 3 1 1 を用いて、ストレージ 4 0 0 中のデータにアクセスを行う。また、I O 管理部 3 1 1 は、ストレージ 4 0 0 に再開可能ポイントを通知するための再開可能ポイント通知部 3 1 2 を有する。また、ホスト 6 0 1 にアプリケーション 6 0 3 a、6 0 3 b の実行イメージを転送し、再開可能ポイント通知部 3 1 2 に再開可能ポイントを通知する実行イメージ転送部 6 0 4 が備えられている。

【 0 2 8 6 】

本実施の形態におけるアプリケーション 6 0 3 a、6 0 3 b は、再開機能を有

さないアプリケーションである。すなわち、ストレージの記憶媒体 1 0 1 のデータ記録状態が所定の状態になっていれば処理を再開できるような機能を有していないアプリケーションである。

【 0 2 8 7 】

図 3 9 は、ホスト 6 0 1 の構成例を示すブロック図である。図 3 9 に示すように、ホスト 6 0 1 には、実行イメージを保存する実行イメージ保存部 6 0 6 と、ホスト 6 0 0 から送られてきた実行イメージを受け取り実行イメージ保存部 6 0 6 に保存する実行イメージ受信部 6 0 5 と、ホスト 6 0 0 の状況を監視するホスト監視部 6 0 8 と、実行イメージ保存部 6 0 6 中の実行イメージを元にアプリケーションを再開させるアプリケーション再開部 6 0 7 とが備えられている。実行イメージ保存部 6 0 6 は、揮発性半導体メモリ、不揮発性半導体メモリ、磁気ディスク、光磁気ディスク、光ディスク等のデータを保存する媒体である。

【 0 2 8 8 】

実行イメージは、稼働系の処理実行状態を示す情報であり、各アプリケーションによって実行されるプロセスを他のホストで実行させるのに必要な情報である。実行イメージには、例えば、各プロセスの仮想アドレス空間中のデータ、プロセス管理情報であるレジスタの値、プログラムカウンタの値およびプロセスの状態等である。プロセスが使用しているファイルやプロセス間通信を復元するための情報等が含まれる。また、並列処理によって一つのプロセスを複数のスレッドで実現する場合には、各スレッドのレジスタの値、プログラムカウンタの値、プログラム状態ワードおよび各種フラグ等も実行イメージに含まれる。さらに、実行イメージは、対象プロセスが使用するカーネル中の通信バッファの内容等を含む場合もある。なお、スレッドとは、並列処理における処理単位である。

【 0 2 8 9 】

次に、動作について説明する。まず、実行イメージ転送時の動作を説明する。正常系のホスト 6 0 0 の実行イメージ転送部 6 0 4 には、例えば、ユーザから実行イメージ転送指示が入力される。この指示が入力されるタイミングは、任意のタイミングでよい。指示が入力されると、実行イメージ転送部 6 0 4 は、アプリケーション 6 0 3 a, 6 0 3 b の実行イメージを取得する。次に、実行イメージ

転送部 6 0 4 は、I O 管理部 3 1 1 の再開可能ポイント通知部 3 1 2 に、ストレージ 4 0 0 に対して再開可能ポイントを通知するように指示する。そして、再開可能ポイント通知部 3 1 2 は、ストレージ 4 0 0 に再開可能ポイントを通知する。次いで、実行イメージ転送部 6 0 4 は、ホスト 6 0 1 の実行イメージ受信部 6 0 5 に対して、取得した実行イメージを転送する。ホスト 6 0 1 において、実行イメージ受信部 6 0 5 は、ホスト 6 0 0 の実行イメージ受信部 6 0 6 から実行イメージを受け取ると、受け取った実行イメージを実行イメージ保存部 6 0 6 に保存する。

【 0 2 9 0 】

ストレージ 4 0 0 は、実行イメージの転送が指示された時点において再開可能ポイント通知を受信し、その通知に応じて動作する。従って、待機系のストレージ 4 0 1 は、正常系のホストにおいて実行イメージ転送が指示された時点におけるスナップショットを作成する。

【 0 2 9 1 】

ストレージとして第 7 の実施の形態と同様のストレージ 3 0 1, 3 0 2 (図 1 9 参照) を用いる場合、待機系のストレージ 4 0 1 は、記憶媒体の状態を、正常系のホストにおいて実行イメージ転送が指示された時点における状態に維持する。

【 0 2 9 2 】

なお、実行イメージの容量は、数百 M B 以上になる場合もあり、ホスト 6 0 1 への実行イメージの転送には時間がかかる。そして、実行イメージの転送中にアプリケーション 6 0 3 a, 6 0 3 b が処理を進めるとメモリやレジスタの値が変化してしまい、転送途中で実行イメージの内容が変わってしまう。そのため、実行イメージ転送中は、アプリケーション 6 0 3 a, 6 0 3 b は処理を停止する。

【 0 2 9 3 】

あるいは、ホスト 6 0 0 が備えるメモリや磁気ディスク等 (図 3 7 において図示せず) に、実行イメージの情報を保存し、保存した実行イメージをホスト 6 0 1 に送信してもよい。一旦、保存完了後に、処理を進めても、保存された実行イメージの内容は保たれる。従って、この場合、ホスト 6 0 0 のアプリケーション

603 a, 603 b は処理を進めることができる。実行イメージをメモリ等に保存する時間は、実行イメージの転送時間に比べ短い。従って、実行イメージをメモリ等に保存してから転送するようにすれば、処理を停止する時間は短くて済む。

【0294】

ここでは、実行イメージ転送部604がユーザからの指示に応じて実行イメージを転送する場合を示した。実行イメージ転送部604が実行イメージ転送処理を開始するタイミングは、ユーザから指示が入力された時点に限定されない。例えば、一定時間間隔毎に実行イメージ転送処理を開始するようにしてもよい。また、実行イメージの転送が終了したときに再度実行イメージの転送処理を開始してもよい。あるいは、アプリケーションがアイドルになった時点、アプリケーションのアイドルタイムが一定値を越えた時点等に転送処理を開始してもよい。

【0295】

次に、災害発生時の動作を説明する。災害等によってホスト600が使用できなくなった場合には、ストレージ401を待機系から正常系にする。また、ホスト601が、ホスト600から処理を引き継ぐ。

【0296】

障害検知からアプリケーション再開までのホスト601の動作を説明する。まず、ホスト601中のホスト監視部608がホスト600の異常を検出する。例えば、ホスト監視部608は、ホスト600と常時あるいは定期的に通信を行い、一定時間ホスト600と通信できなくなった場合に異常が生じたと認識する。あるいは、ホスト600から異常が報告された場合に、ホスト600において災害が発生したと認識してもよい。

【0297】

ホスト監視部608は、ホスト600の異常を検出すると、待機系であるストレージ401に対してスナップショット復帰コマンドを発行する。そして、スナップショット復帰コマンドに対する応答を待ち、応答を受けたら、ホスト監視部608は、アプリケーション再開部607に、アプリケーションの再開を指示する。アプリケーション再開部607は、実行イメージ保存部606に保存されて

いる実行イメージを用いてアプリケーションアプリケーションの動作を再開させる。本例では、ストレージとして第五の実施の形態と同様のストレージ400, 401を用いる場合を示した。ストレージとして第7の実施の形態と同様のストレージ301, 302（図19参照）を用いる場合、ホスト監視部608は、スナップショット復帰コマンドの代わりに遅延データ破棄コマンドを出力すればよい。

【0298】

なお、アプリケーション再開部607は、アプリケーションの処理を再開するときに、実行イメージのうち動作するホストによって変更を必要とする情報がある場合には、復元時にその情報を変更する。

【0299】

第9の実施の形態では、正常系のホストにおいて実行イメージの転送が指示された時点で、正常系のストレージは再開可能ポイント通知を受信する。従って、待機系のストレージは、正常系のホストにおいて実行イメージの転送が指示された時点におけるスナップショットを作成する。あるいは、正常系のホストにおいて実行イメージの転送が指示された時点における記憶媒体の状態を維持する。さらに、待機系のホストは、実行イメージの転送が指示された時点における正常系ホストの実行イメージを保持する。従って、アプリケーションによって再開機能が実現されていなくても、待機系のストレージは、所定の時点における正常系のストレージの状態を再現でき、待機系のホストは、その時点における実行イメージを再現できるので、待機系で処理を迅速に再開することができる。

【0300】

第9の実施の形態において、実行イメージ転送部604は、実行イメージを全て転送するのではなく、前回転送した実行イメージとの差分のみを転送してもよい。このように、前回転送した実行イメージから変更された部分のみを転送するようにすれば、転送時間を短縮化することができる。

【0301】

仮想記憶をコンピュータシステムでは、使用される確率の高い論理アドレスと物理アドレスとの変換テーブルを用いることが多い。この変換テーブルは、TL

B (Translation Look-aside Buffer) と呼ばれる。TLBでは、変換される各アドレス毎に、そのアドレスのデータが書き換えられたか否かを示す更新情報を管理する。この更新情報は、一般に、ダーティ (dirty) フラグまたはダーティビットと呼ばれている。例えば、「コンピュータ・アーキテクチャ 設計・実現・評価の定量的アプローチ (デイビット・パターソン, ジョン・ヘネシー著、富田眞治 他2名訳、日経BP社, 1994年2月, p. 442-443)」では、この更新情報はダーティビットとして記載されている。実行イメージ転送部604は、変更があったことを示すダーティフラグに対応する情報を実行イメージとして送信すればよい。

【0302】

また、仮想記憶において、プログラムの実行に必要なページを実記憶装置に配置することをページインとよび、ページを実記憶装置から外して仮想記憶装置に配置することをページアウトとよぶ。実行イメージ転送部604は、実行イメージの転送を指示された場合には、ページインしているメモリのデータをホスト601に送信し、その後、ページアウトが発生したタイミングで、ページアウトしたメモリのデータをホスト601に送信してもよい。ページアウトしたメモリのデータは、次にページインするまで変更されることがない。そのため、このように実行イメージを送信することで、実行イメージの送信量を減らすことができる。

【0303】

第9の実施の形態において、実行イメージ転送手段は、実行イメージ転送部604によって実現される。実行イメージ保存手段は、実行イメージ保存部606によって実現される。任意時点再開可能ポイント通知手段は、ホスト600の再開可能ポイント通知部312によって実現される。

【0304】

図40, 41は、クライアントサーバシステムに、第7の実施の形態から第9の実施の形態を適用した場合の構成例を示すブロック図である。通常、ホスト300が各クライアント500-1~500-nのサーバとして機能する。災害等により、ホスト300またはストレージ400に以上が発生した場合、ホスト3

03が稼働する。ホスト303が稼働を開始した後、各クライアントはホスト303をサーバとして、ホスト303に処理を要求する。

【0305】

なお、図40に示すように、ホスト300とホスト303とが直接接続されている場合には、第7の実施の形態から第9の実施の形態で説明したように、ホスト303がホスト300の状態を監視する。図41に示すように、ホスト300とホスト303とが直接接続されていない場合には、各クライアント500-1～500-nがホスト300の状態を監視し、ホスト300に以上が発生したときに、ホスト303に以上の発生を通知してもよい。

【0306】

第7の実施の形態から第9の実施の形態において、第1の実施の形態から第4の実施の形態と同様に、中継装置を介して待機系のストレージにデータを送信するようにしてもよい。その場合、第1の実施の形態から第4の実施の形態の効果も得ることができる。また、ストレージ同士で、あるいはストレージと中継装置との間でデータを送受信するときには、第5の実施の形態または第6の実施の形態に示したように冗長化されたデータ群を送受信するようにしてもよい。その場合、第5の実施の形態または第6の実施の形態と同様の効果も得られる。

【0307】

【発明の効果】

本発明によれば、第一のストレージから第二のストレージに転送されるデータを中継する中継装置であって、第一のストレージが災害によって稼働できない状態になっても、稼働を継続できるとあらかじめ算定された位置に設置された中継装置を備え、第一のストレージは、データ転送の制御を行うデータ転送処理手段を含み、データ転送処理手段は、中継装置に対してデータ転送を完了したときに、第二のストレージに対してデータ転送を完了したとみなす。従って、稼働系のストレージが被災しても中継装置は被災せず、また、稼働系の上位装置は、待機系のストレージにデータが格納されるのを待たずに、次の処理を開始することができるので、データ転送に伴う処理の遅れが改善される。

【0308】

また、本発明によれば、送信される元データから少なくとも1つのエラー訂正のための冗長データを作成し、元データと冗長データとを別々のデータ送信単位で送信する。従って、待機系では、元データと冗長データとの集合のうちの一部から元データを復元することができ、送信過程で一部のデータが廃棄されても再度送信する必要がない。その結果、データの転送を迅速に完了させることができる。

【0309】

また、本発明によれば、稼働系のストレージは、データの書き込み要求が発生すると、書き込み対象のデータと遅延書き込み要求とを待機系のストレージに送信する遅延書き込み要求手段と、そのデータの状態であればアプリケーションがそのまま動作を再開可能な再開可能ポイントであることを知らせるための再開可能ポイント通知を上位装置から受けると遅延書き込み実行要求を待機系のストレージに送信する書き込み実行要求手段とを含み、待機系のストレージは、データを一時的に記憶する一時記憶手段と、受信したデータを遅延書き込み要求に応じて一時記憶手段に格納するとともに、遅延書き込み実行要求を受信すると一時記憶手段に格納されているデータを記憶媒体に格納する格納処理手段とを含む。従って、待機系の記憶媒体の状態を再開可能ポイント通知時の状態にすることができ、正常系で異常が発生したときに待機系ですぐに処理を再開することができる。

【0310】

また、本発明によれば、稼働系のストレージは、データの書き込み要求が発生すると、書き込み対象のデータと書き込み要求とを待機系のストレージに送信する書き込み要求手段と、そのデータの状態であればアプリケーションがそのまま動作を再開可能な再開可能ポイントであることを知らせるための再開可能ポイント通知を上位装置から受けるとスナップショット作成要求を待機系のストレージに送信するスナップショット作成要求手段とを含み、待機系のストレージは、書き込み要求を受信したら書き込み要求に対応するデータを書き込むべき領域を割り当てて記憶媒体にデータを格納し、スナップショット作成要求を受信したらスナップショットを作成するスナップショット作成手段を含む。従って、待機系の

記憶媒体の状態を再開可能ポイント通知時の状態にすることができ、正常系で異常が発生したときに待機系ですぐに処理を再開することができる。

【図面の簡単な説明】

- 【図 1】 データ複製システムの第 1 の実施の形態を示すブロック図である。
- 【図 2】 ストレージの構成例を示すブロック図である。
- 【図 3】 中継装置の構成例を示すブロック図である。
- 【図 4】 処理シーケンサの動作を示すフローチャートである。
- 【図 5】 中継処理部の動作を示すフローチャートである。
- 【図 6】 第 2 の実施の形態の中継処理部の動作を示すフローチャートである。
- 【図 7】 第 3 の実施の形態の処理シーケンサの動作を示すフローチャートである。
- 【図 8】 データ複製システムの第 4 の実施の形態を示すブロック図である。
- 【図 9】 処理シーケンサの動作を示すフローチャートである。
- 【図 1 0】 ストレージの動作を示すフローチャートである。
- 【図 1 1】 ステップ S 2 2 1 の処理を具体的に示すフローチャートである。
- 【図 1 2】 データ転送の例を示すタイミング図である。
- 【図 1 3】 データ複製システムの第 5 の実施の形態を示すブロック図である。
- 【図 1 4】 ストレージの構成例を示すブロック図である。
- 【図 1 5】 処理シーケンサの動作を示すフローチャートである。
- 【図 1 6】 通信部の動作を示すフローチャートである。
- 【図 1 7】 データ複製システムの他の構成例を示すブロック図である。
- 【図 1 8】 第 6 の実施の形態の処理シーケンサの動作を示すフローチャートである。
- 【図 1 9】 データ複製システムの第 7 の実施の形態を示すブロック図である。
- 【図 2 0】 ホストの構成例を示すブロック図である。
- 【図 2 1】 ストレージの構成例を示すブロック図である。
- 【図 2 2】 同期 I D および発行 I D の例を示す説明図である。

【図 2 3】 処理シーケンサの動作を示すフローチャートである。

【図 2 4】 処理シーケンサが遅延write コマンドを受信したときの動作を示すフローチャートである。

【図 2 5】 処理シーケンサの動作を示すフローチャートである。

【図 2 6】 遅延データ反映処理の動作を示すフローチャートである。

【図 2 7】 遅延データ破棄処理を示すフローチャートである。

【図 2 8】 データ複製システムの第 8 の実施の形態を示すブロック図である。

【図 2 9】 ストレージの構成例を示すブロック図である。

【図 3 0】 アドレステーブルの一例を示す説明図である。

【図 3 1】 エントリの説明図である。

【図 3 2】 処理シーケンサの動作を示すフローチャートである。

【図 3 3】 処理シーケンサの動作を示すフローチャートである。

【図 3 4】 L B A 管理部の動作を示すフローチャートである。

【図 3 5】 処理シーケンサの動作を示すフローチャートである。

【図 3 6】 処理シーケンサの動作を示すフローチャートである。

【図 3 7】 データ複製システムの第 9 の実施の形態を示すブロック図である。

【図 3 8】 ホストの構成例を示すブロック図である。

【図 3 9】 ホストの構成例を示すブロック図である。

【図 4 0】 クライアントサーバシステムに本発明を適用した場合の構成例を示すブロック図である。

【図 4 1】 クライアントサーバシステムに本発明を適用した場合の構成例を示すブロック図である。

【符号の説明】

1 0 ホスト

1 1, 1 2 ストレージ

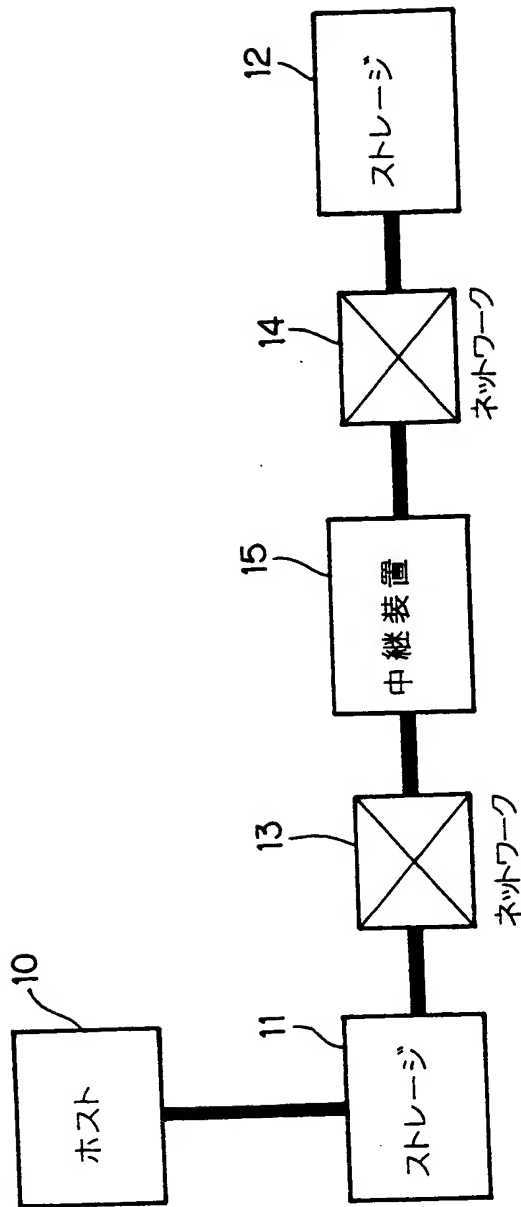
1 3, 1 4 ネットワーク

1 5 中継装置

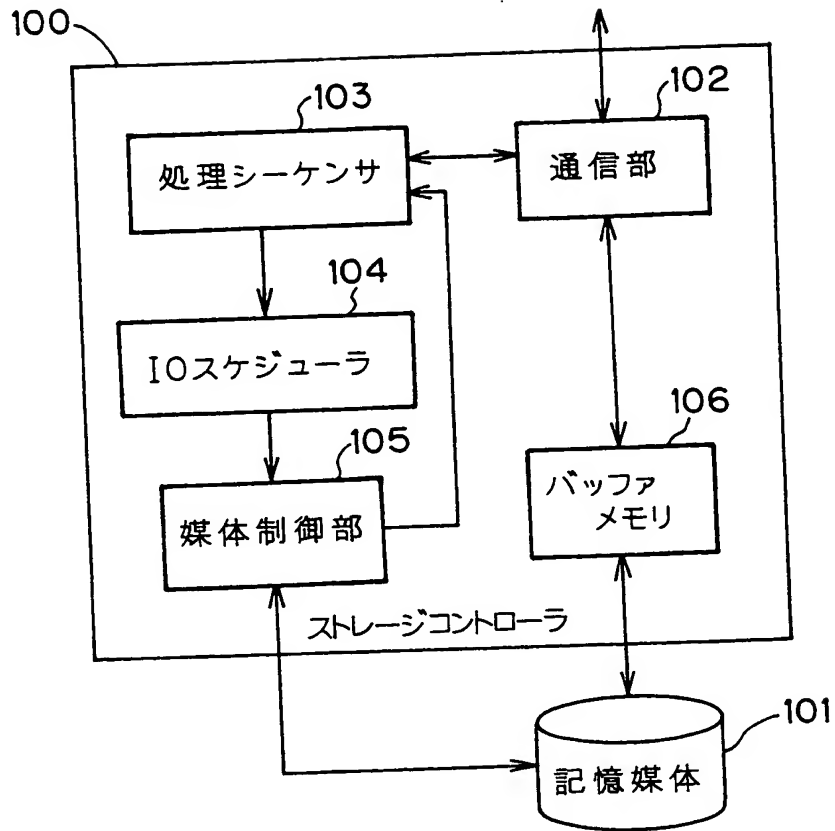
- 1 0 1 記憶媒体
- 1 0 2 通信部
- 1 0 3 処理シーケンサ
- 1 0 4 I Oスケジューラ
- 1 0 5 媒体制御部
- 1 0 6 バッファメモリ
- 1 5 1 中継処理部
- 1 5 0 通信部
- 1 5 2 バッファメモリ

【書類名】 図面

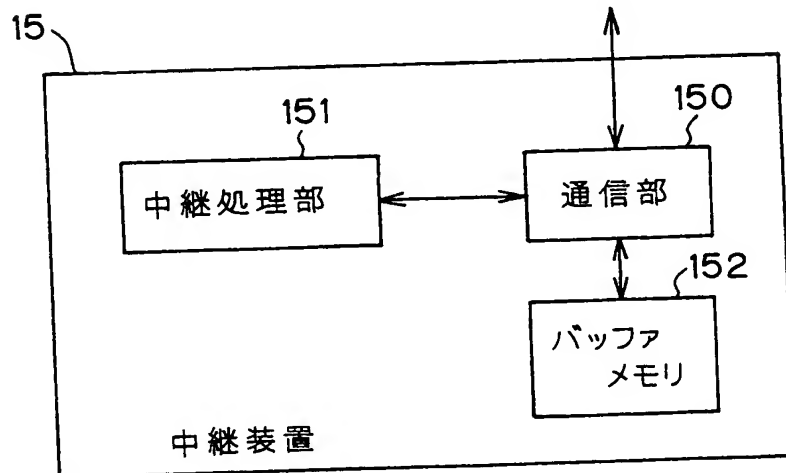
【図1】



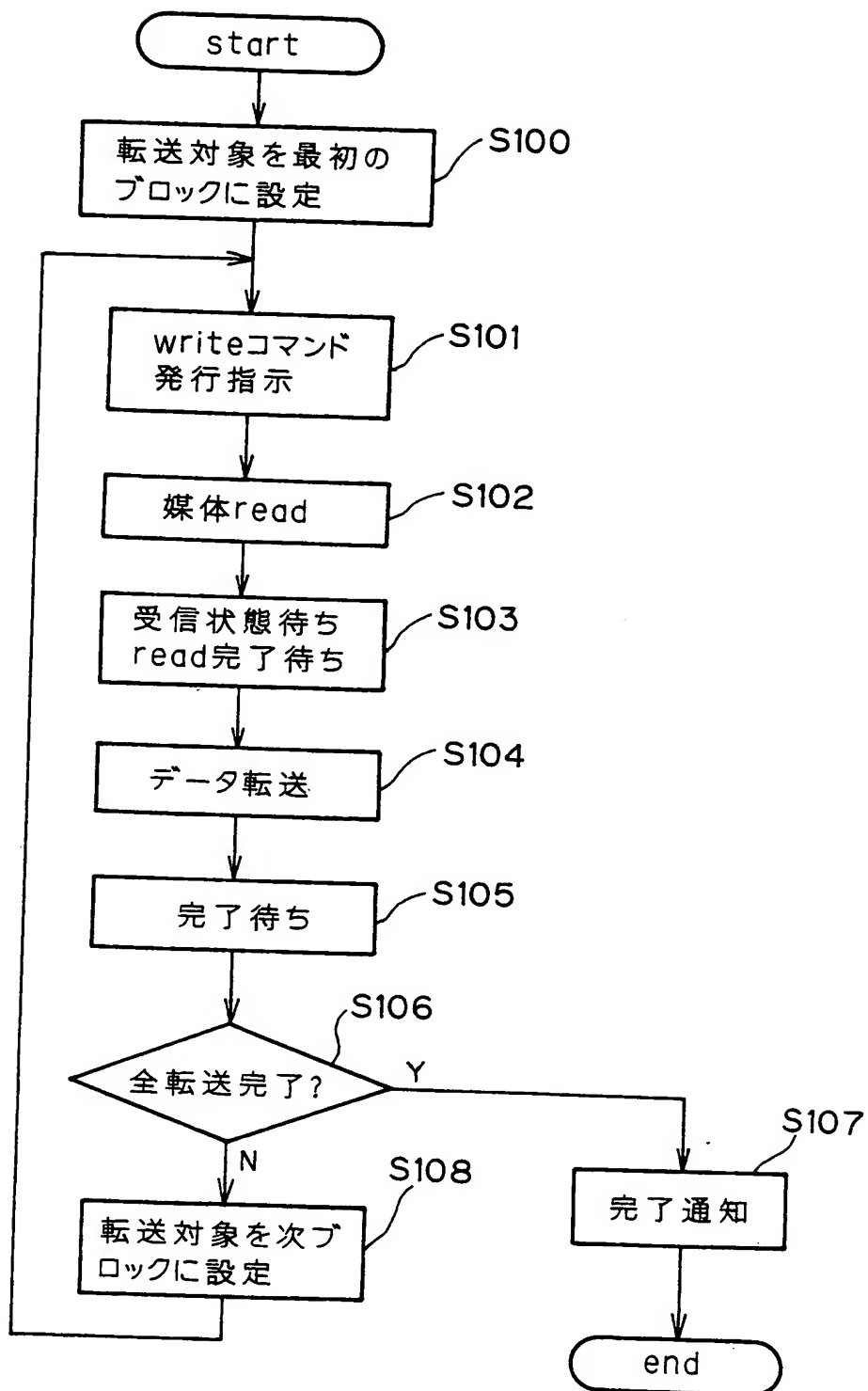
【図2】



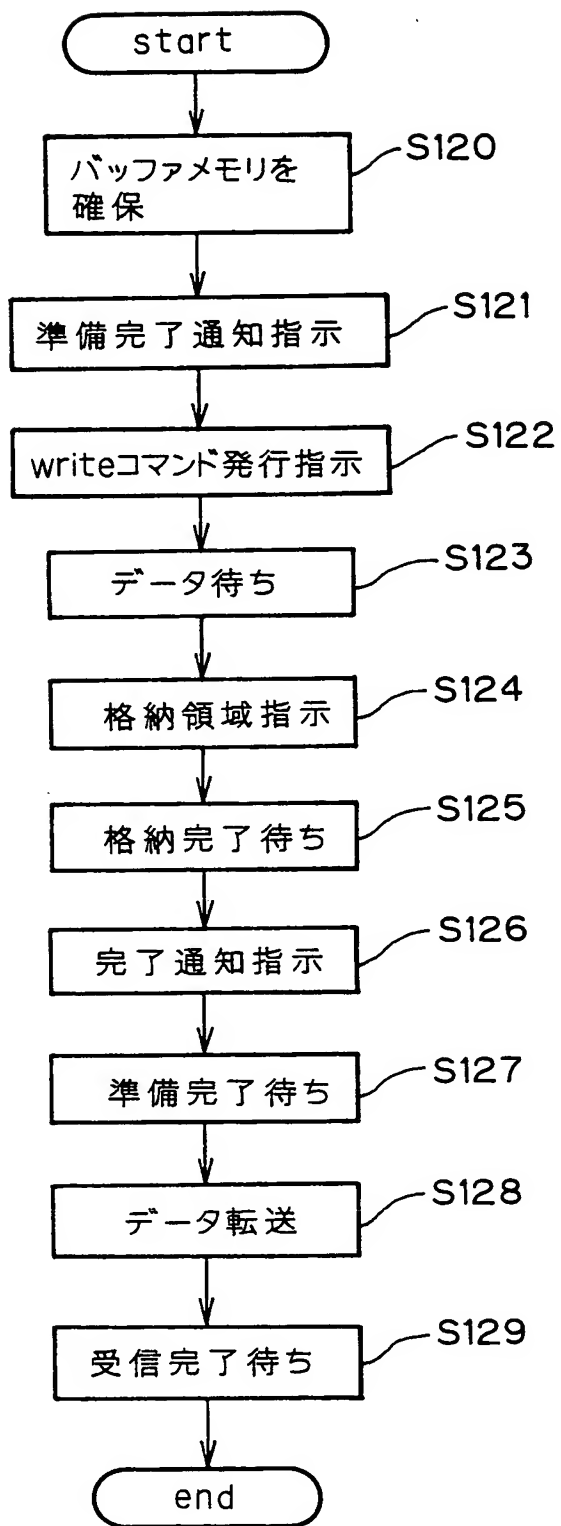
【図3】



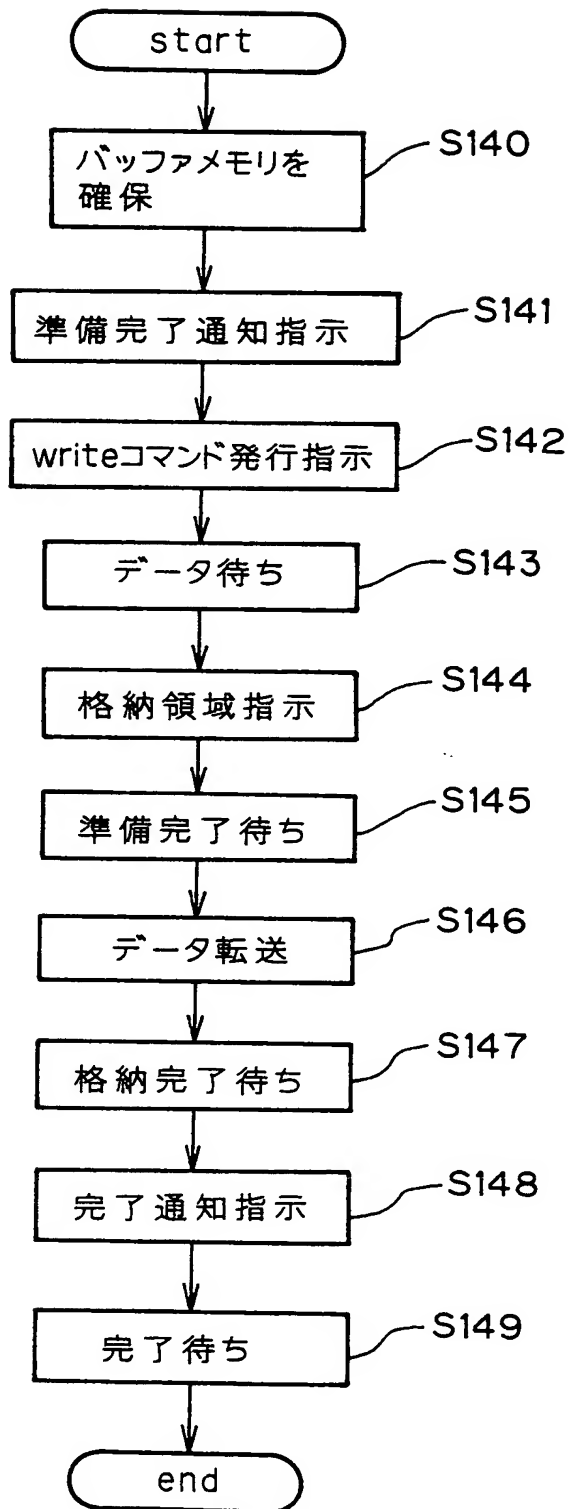
【図4】



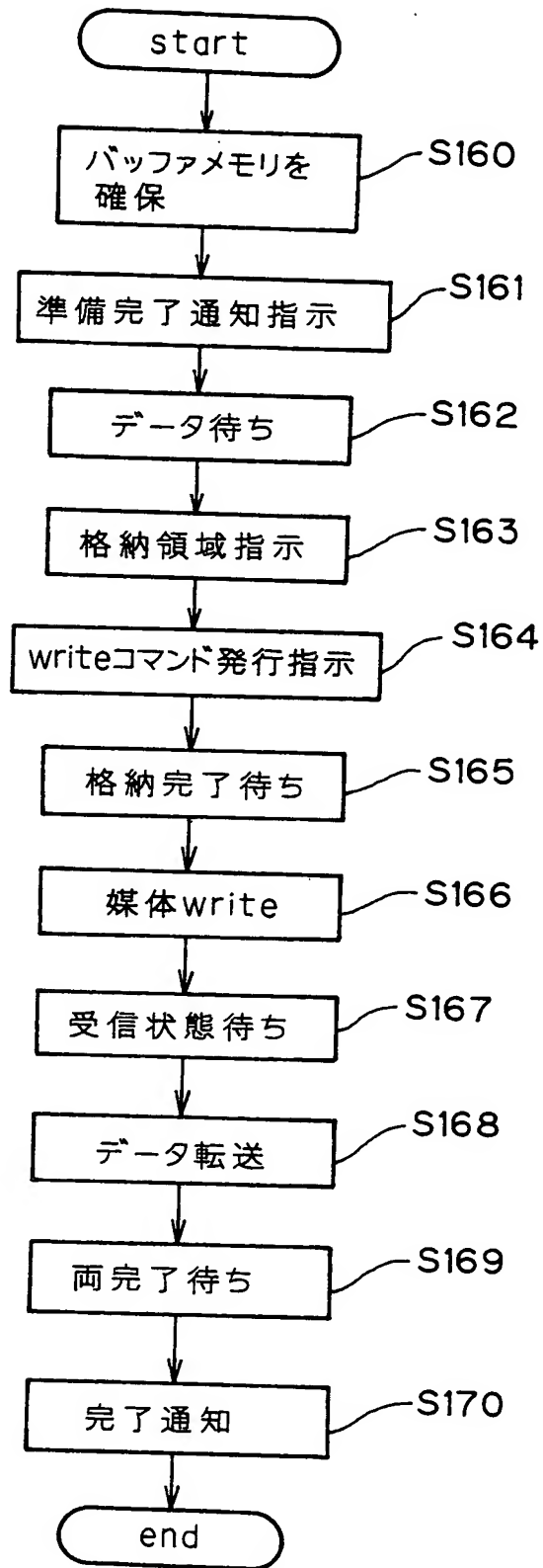
【図 5】



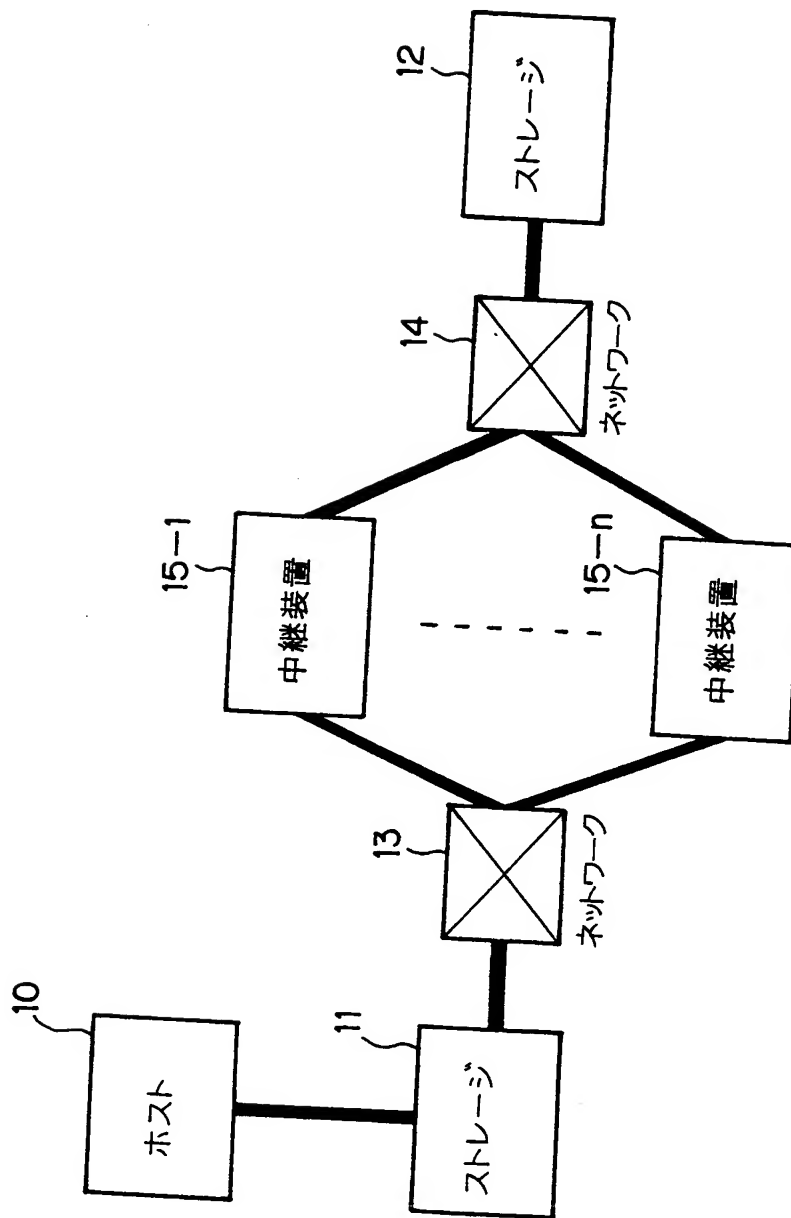
【図 6】



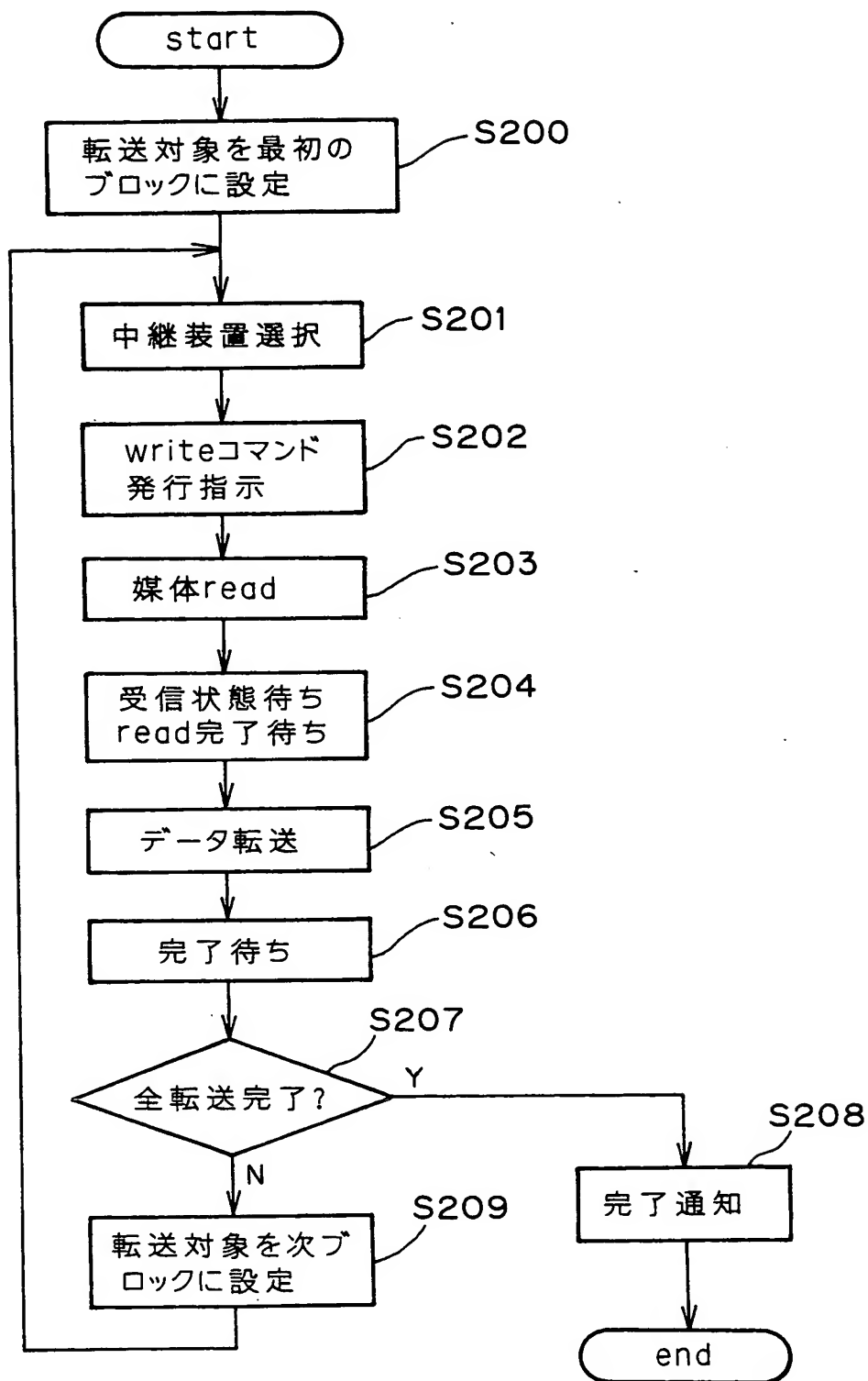
【図 7】



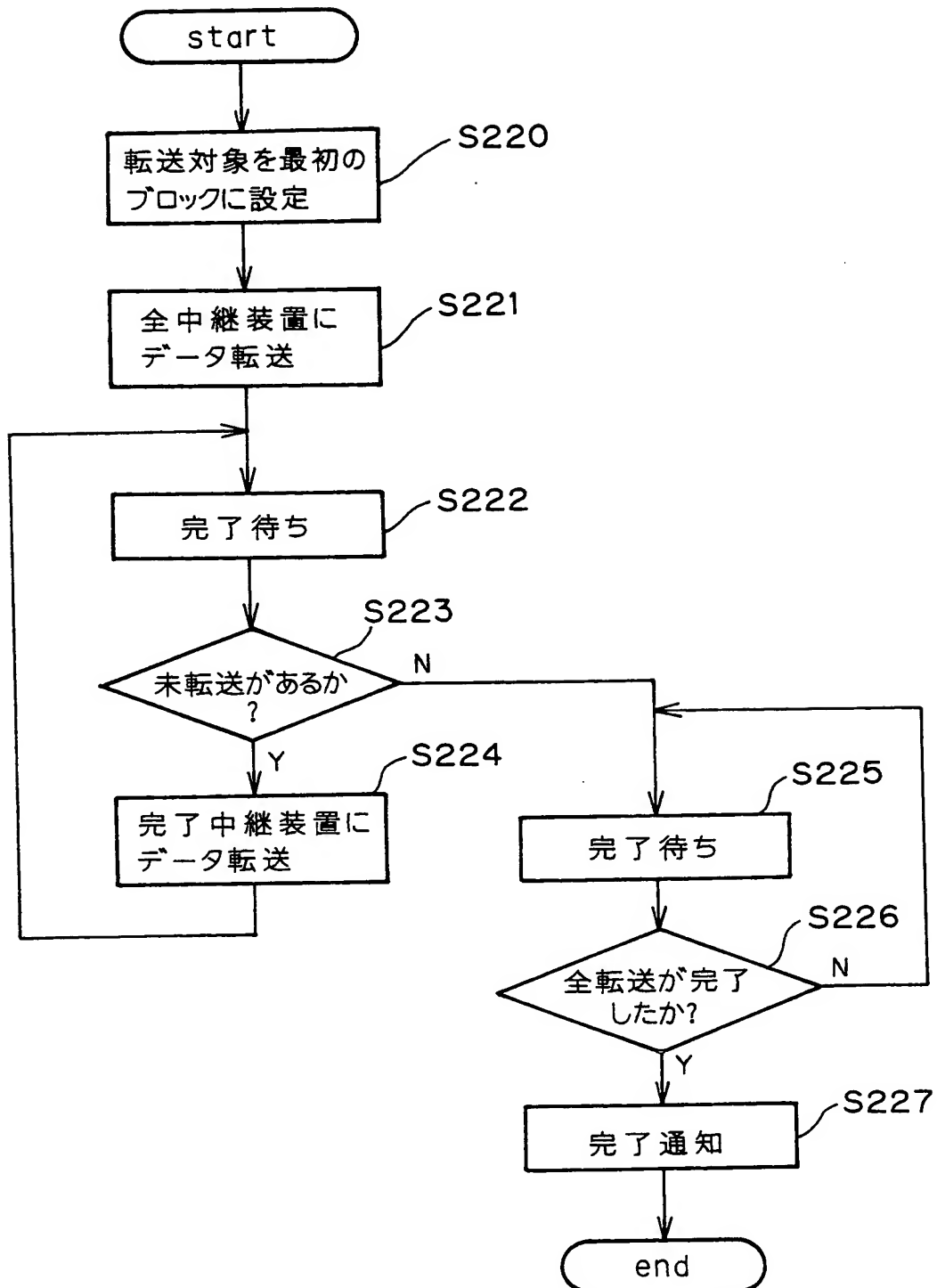
【図8】



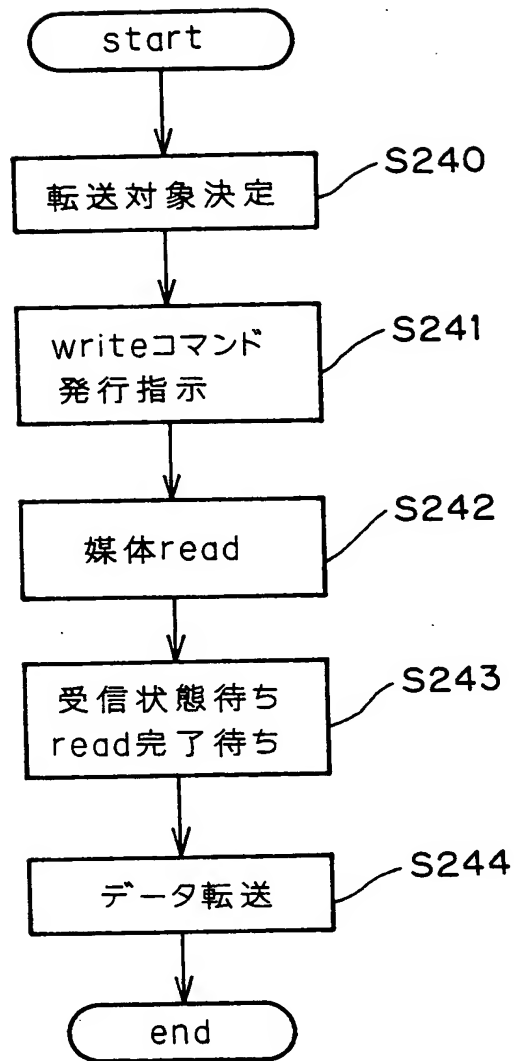
【図 9】



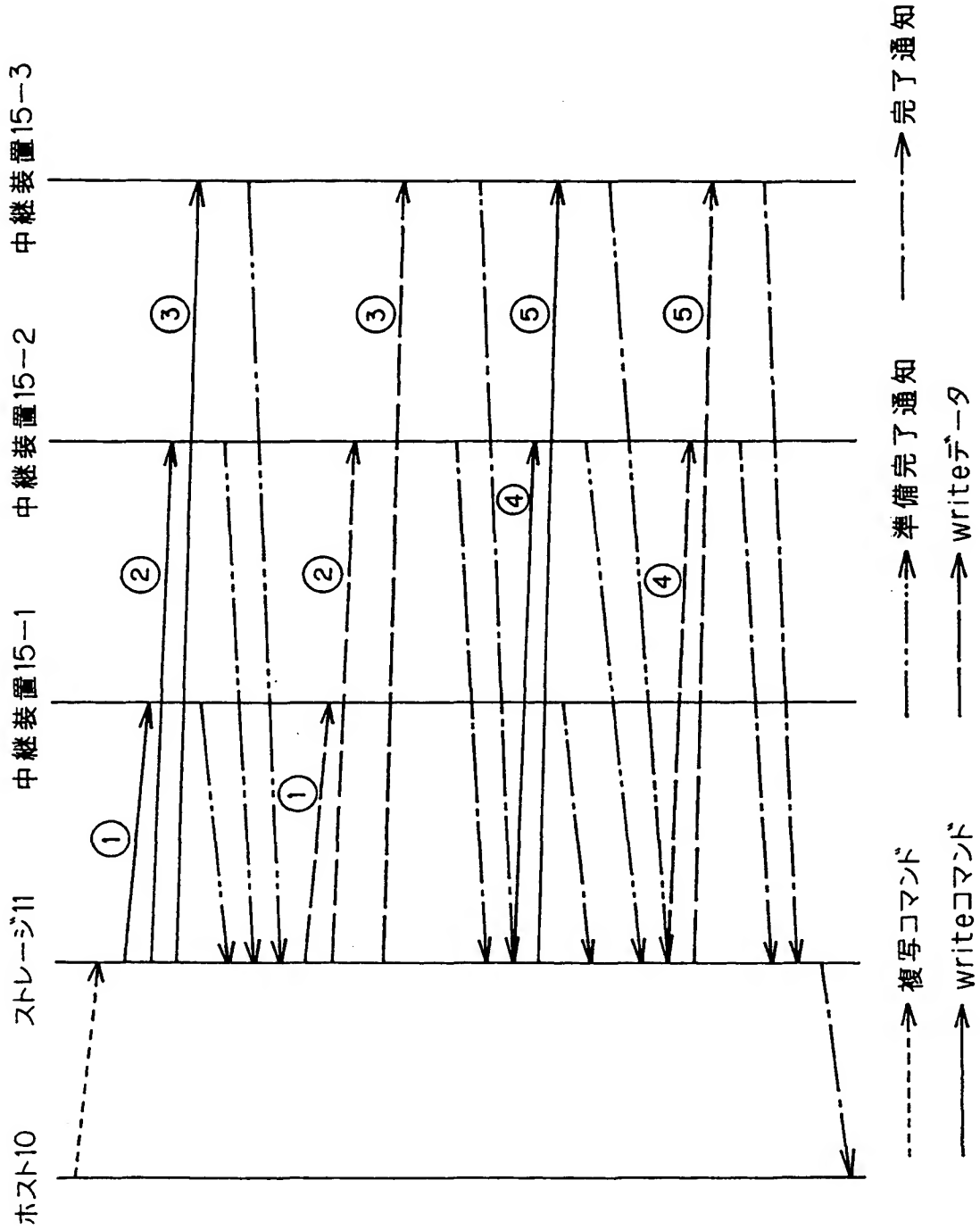
【図10】



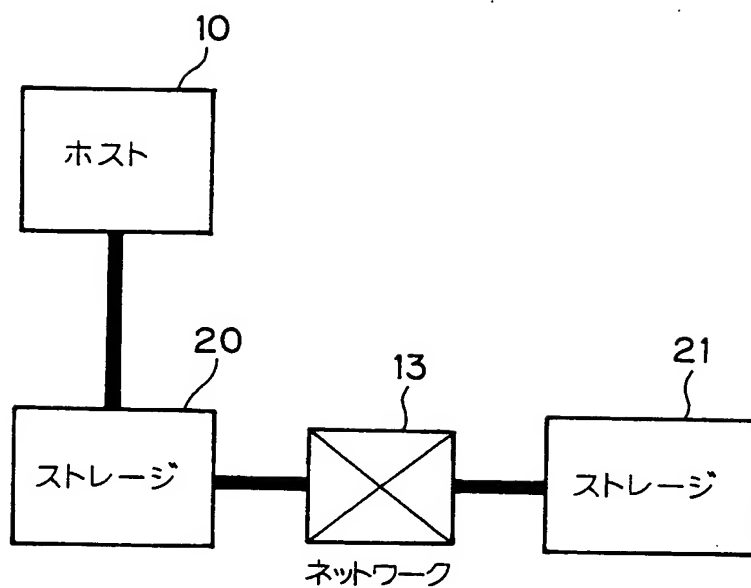
【図11】



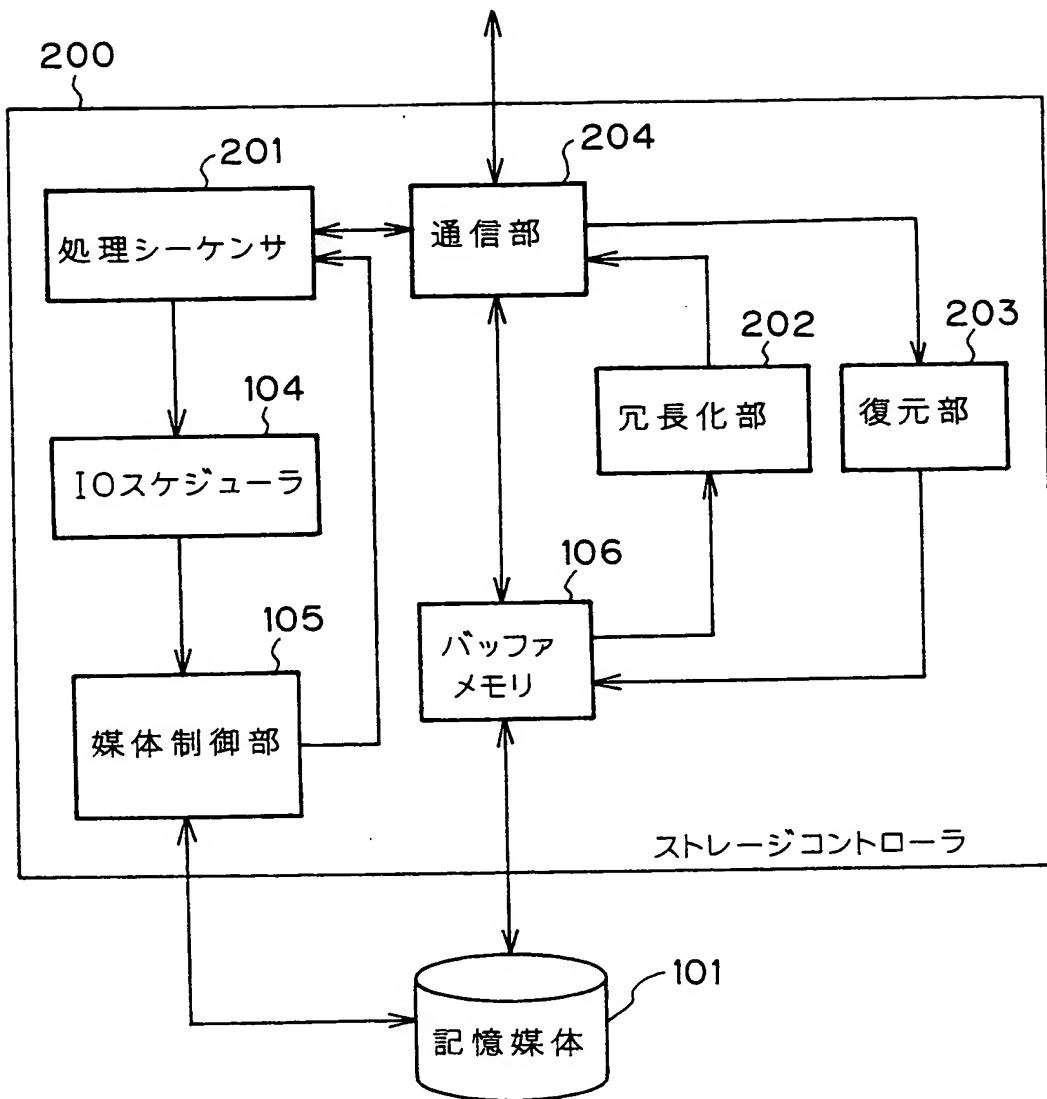
【図12】



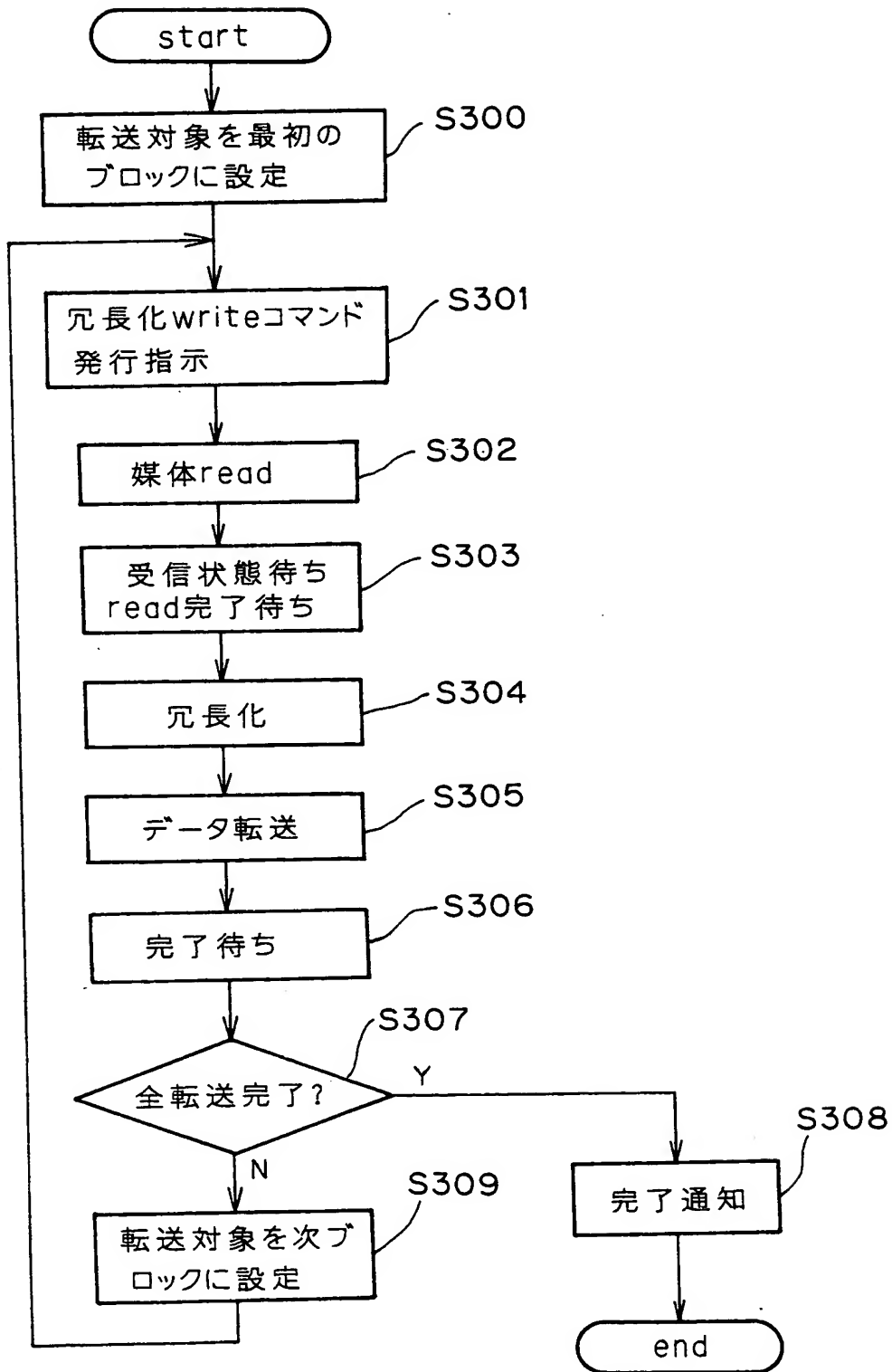
【図 1 3】



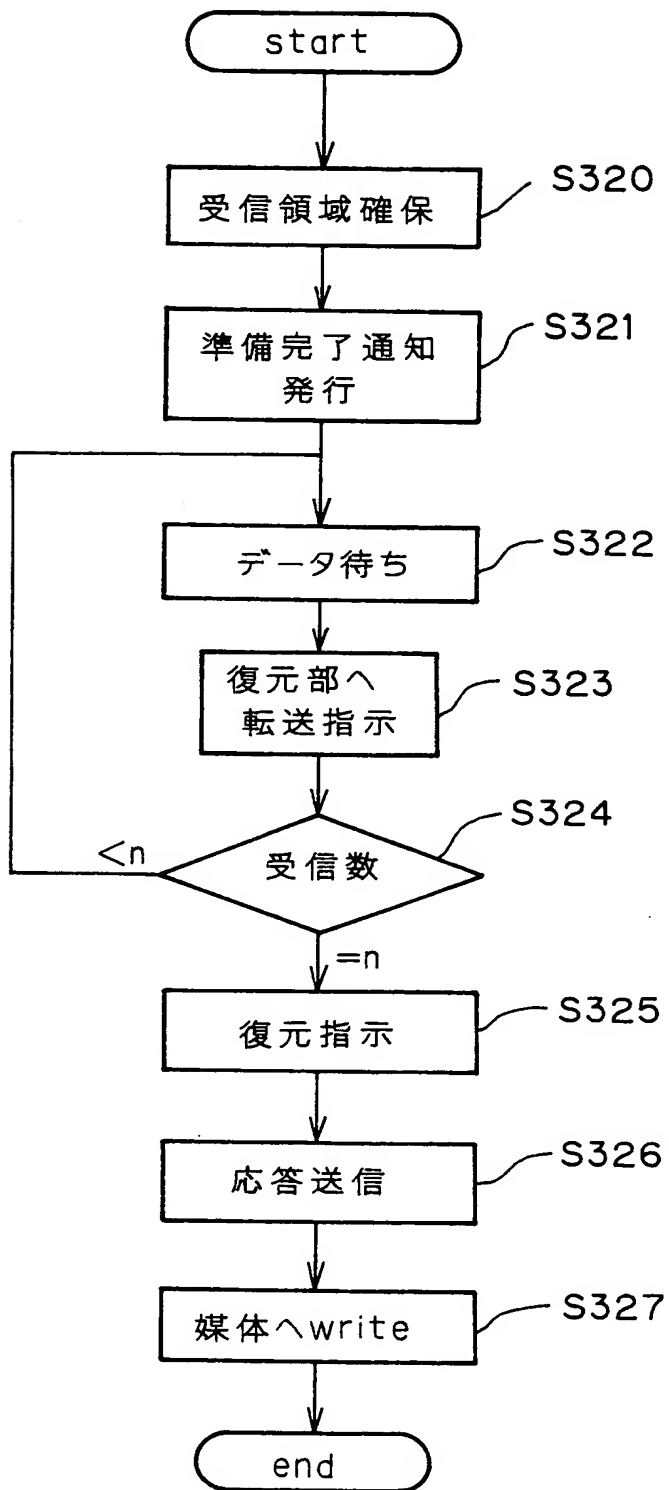
【図 14】



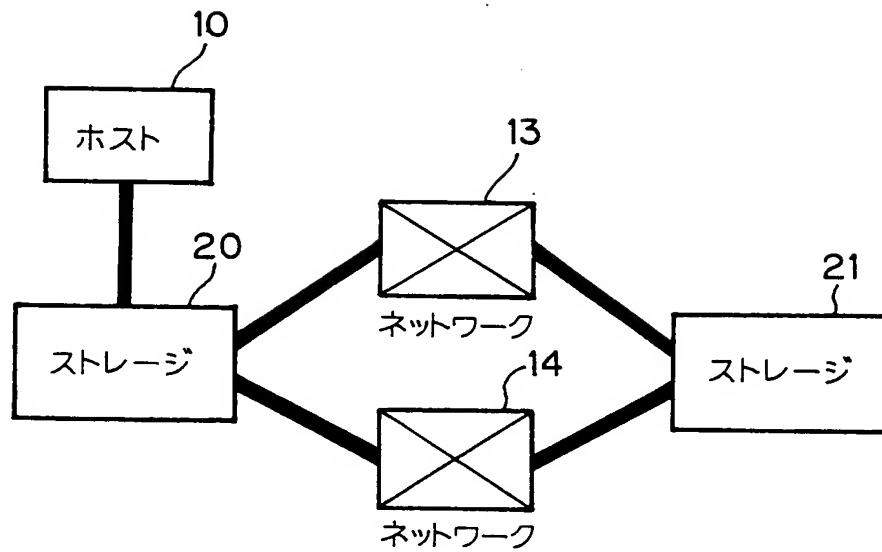
【図15】



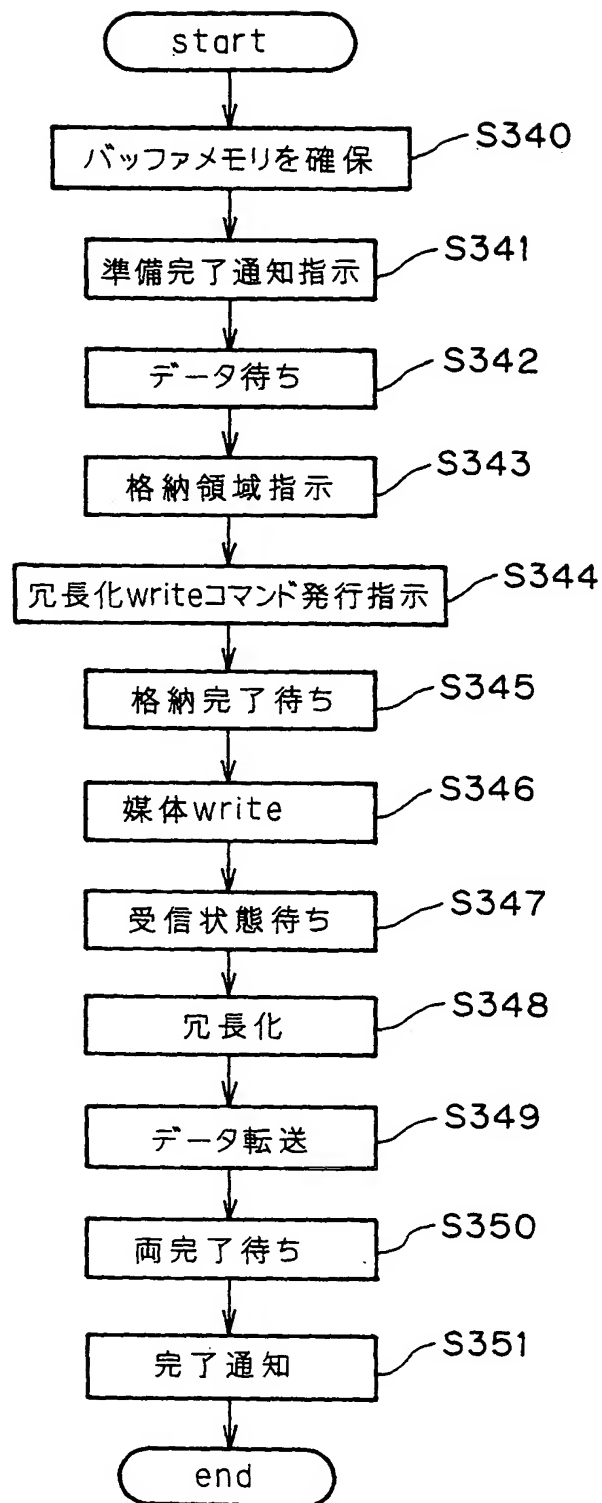
【図 1 6】



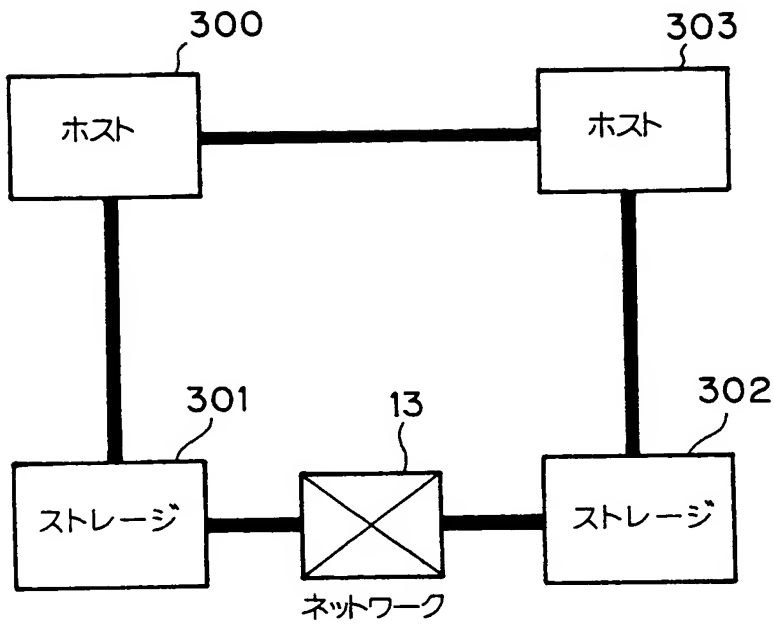
【図 1 7】



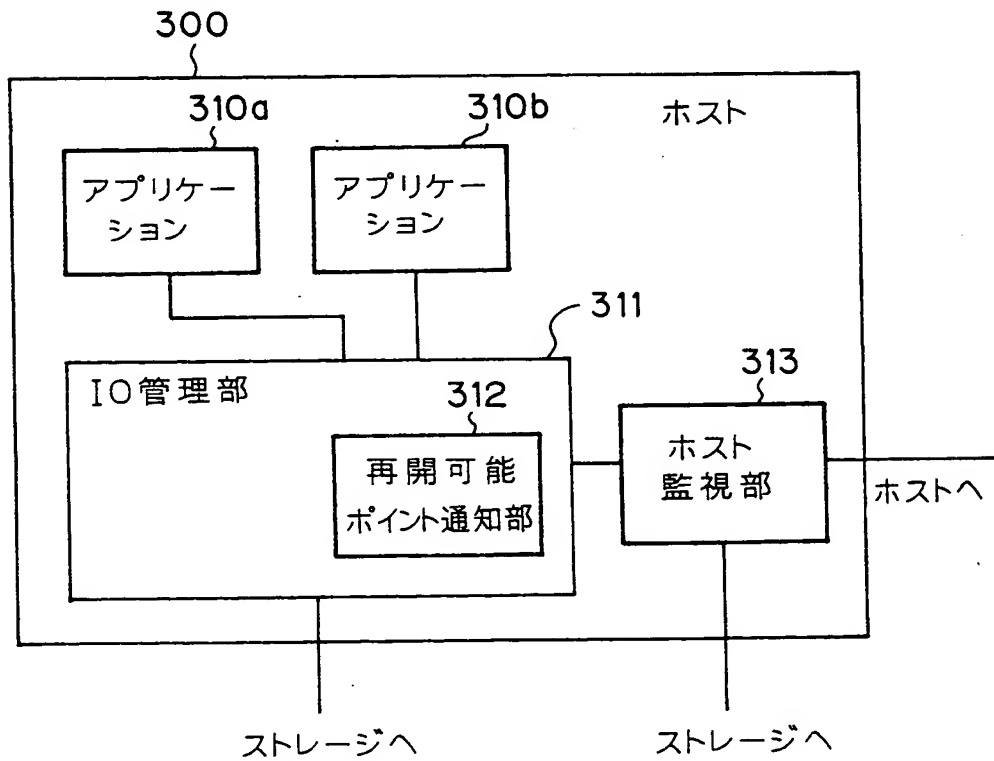
【図18】



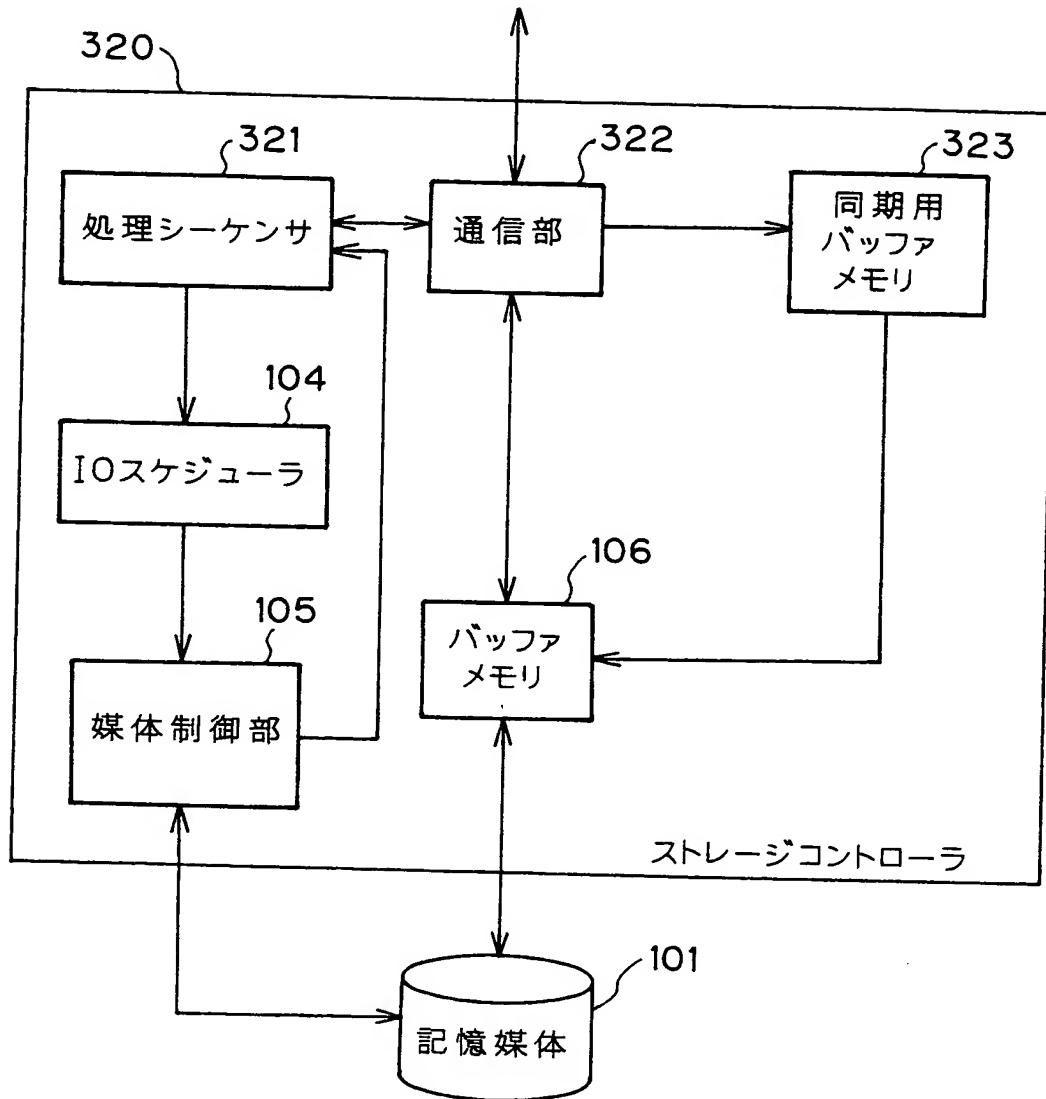
【図19】



【図20】



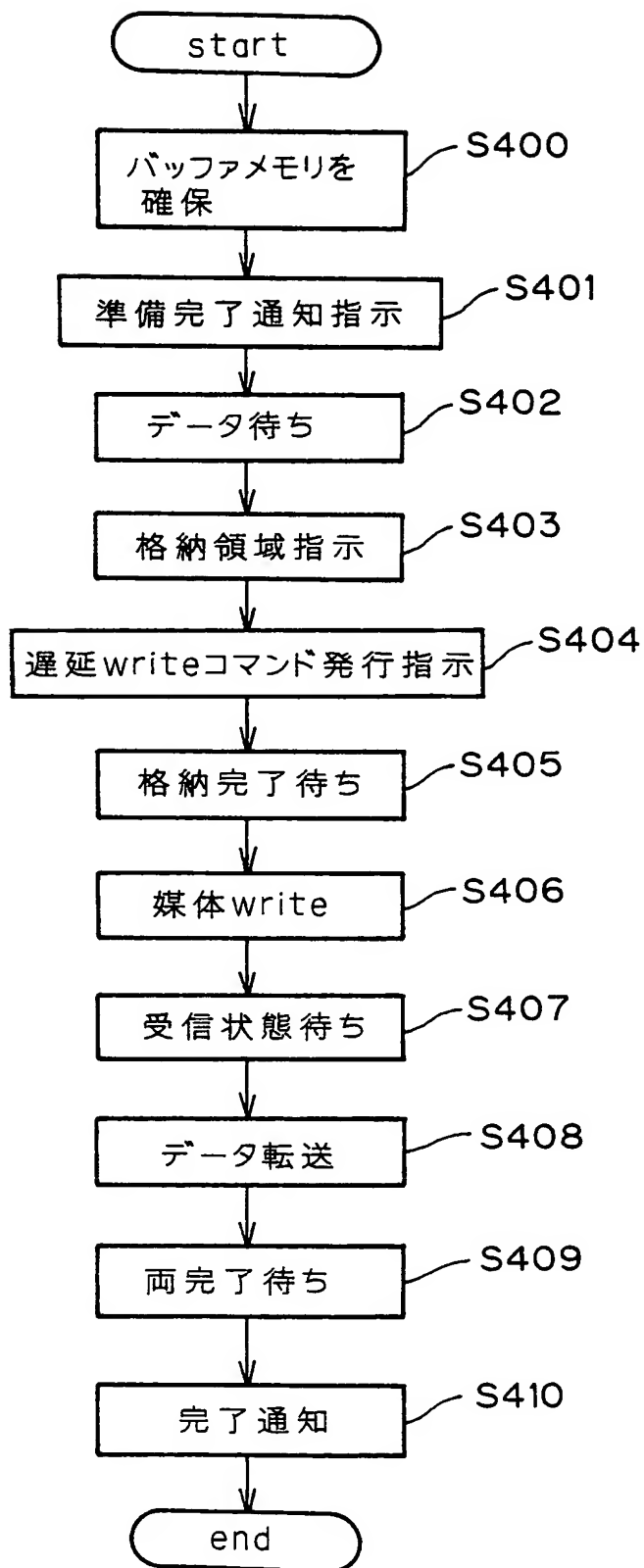
【図 21】



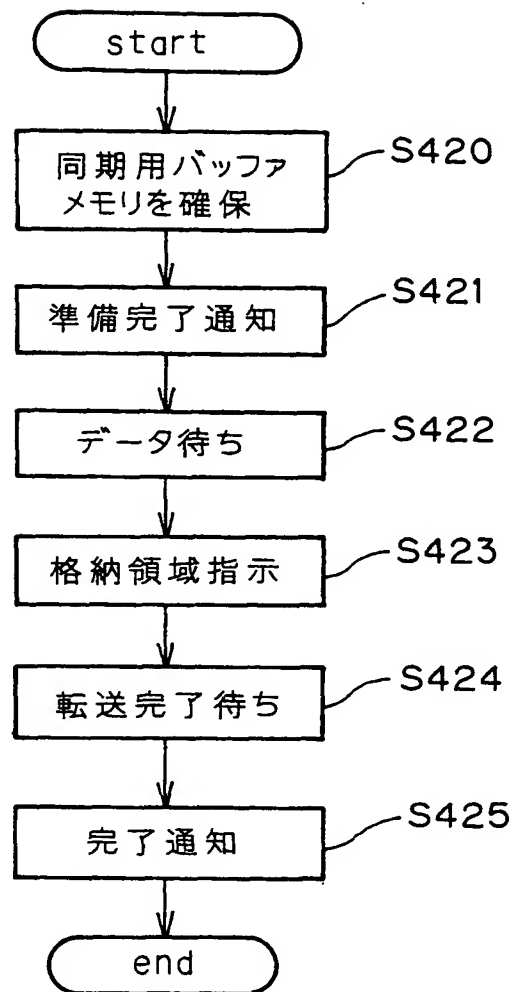
【図 2 2】

再開可能 ポイント (同期ID更新)		再開可能 ポイント (同期ID更新)		再開可能 ポイント (同期ID更新)		再開可能 ポイント (同期ID更新)		再開可能 ポイント (同期ID更新)	
遅延データ 反映コメント	遅延write コメント	遅延データ 反映コメント	遅延write コメント	遅延データ 反映コメント	遅延write コメント	遅延データ 反映コメント	遅延write コメント	遅延データ 反映コメント	遅延write コメント
23	24	24	24	24	24	24	24	25	25
71	72	73	74	75	76	77	78	79	80
同期ID		同期ID		同期ID		同期ID		同期ID	
発行ID		発行ID		発行ID		発行ID		発行ID	
発行順									

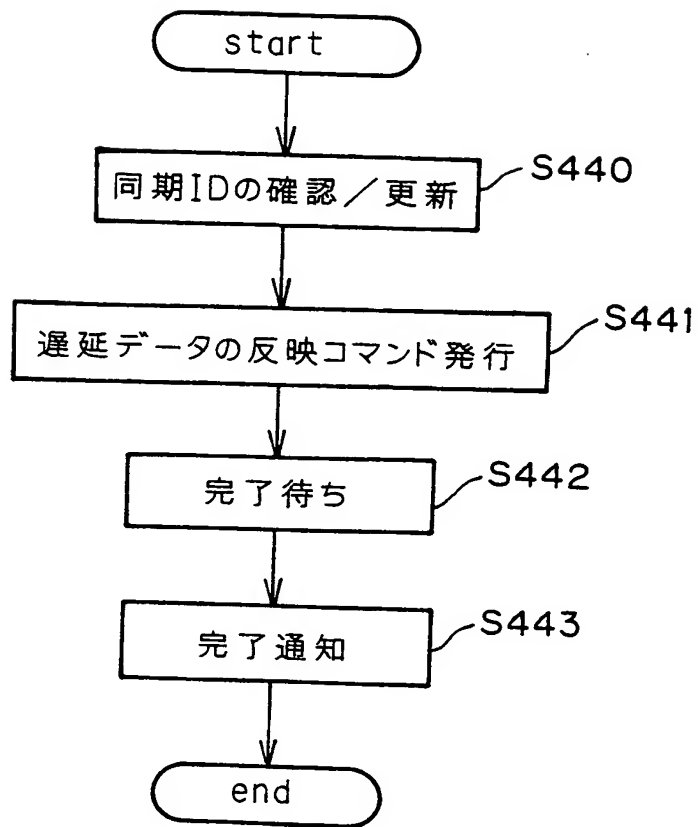
【図 2 3】



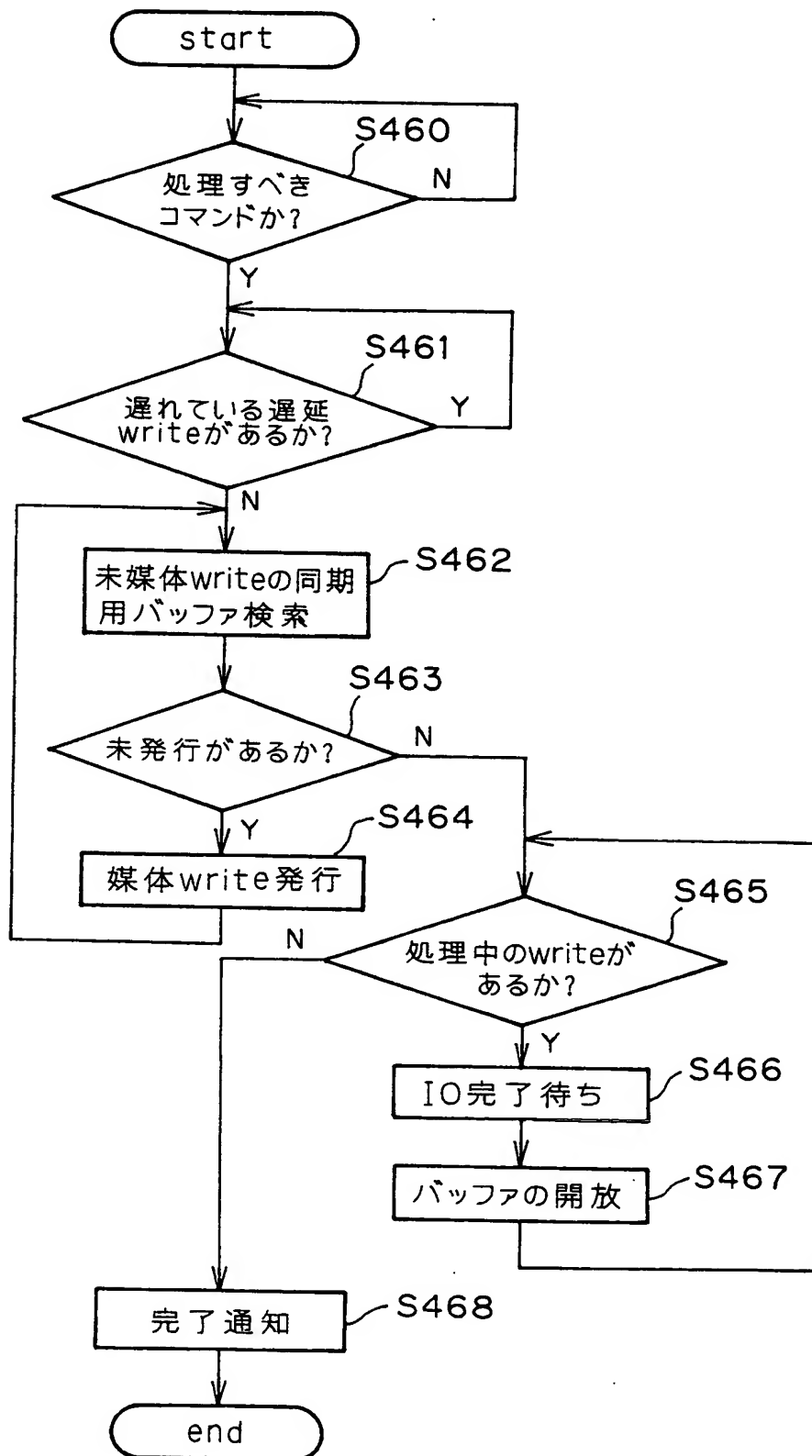
【図 24】



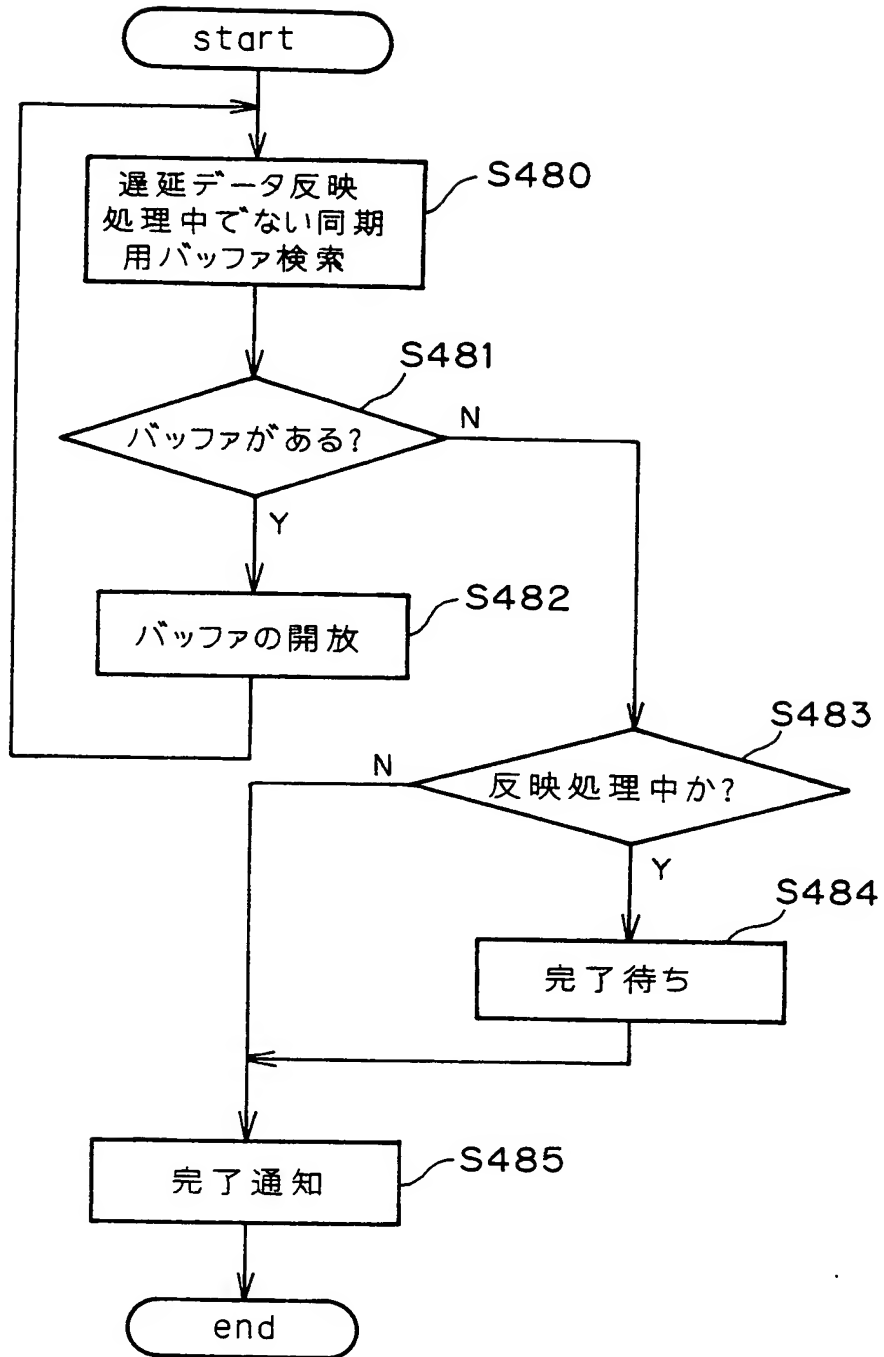
【図 2 5】



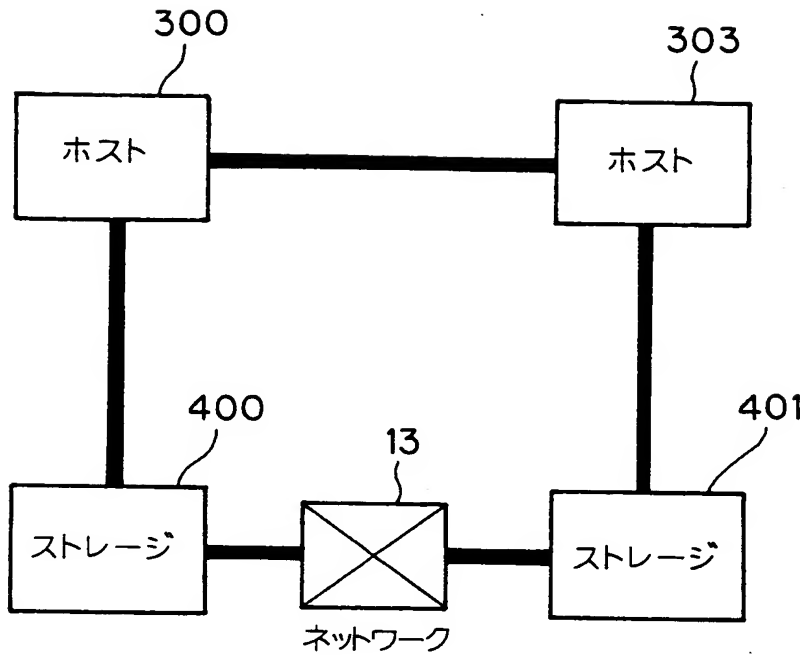
【図 26】



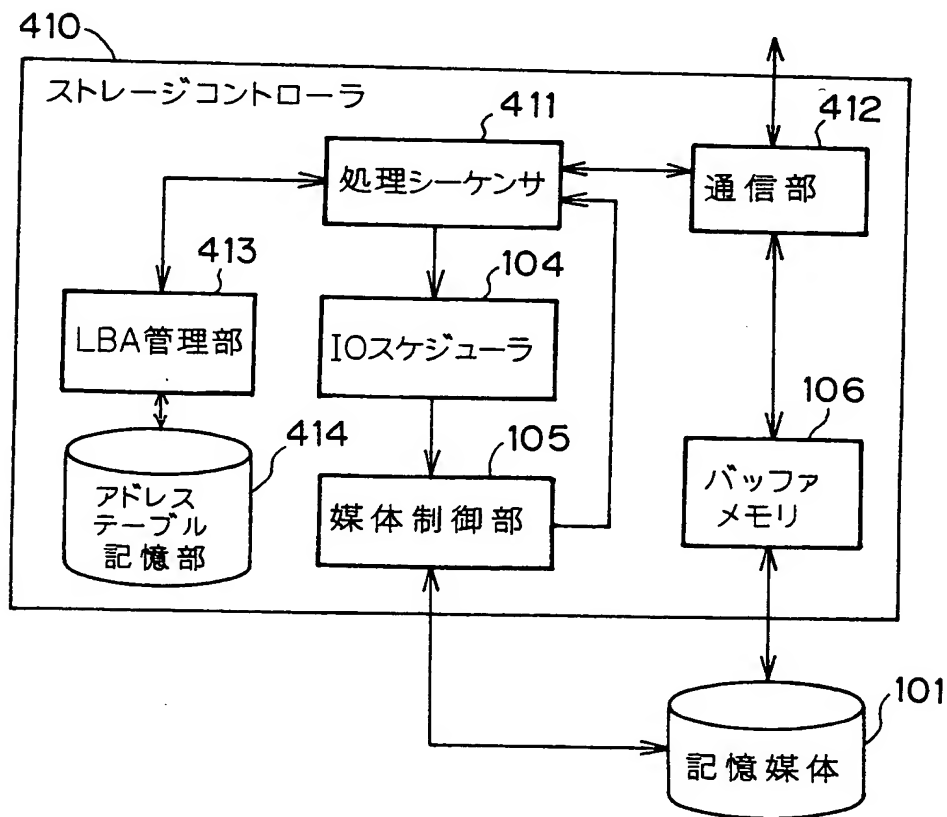
【図 27】



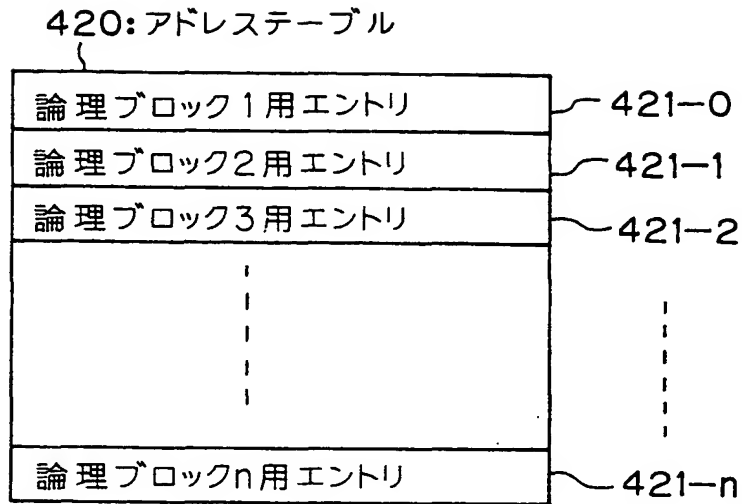
【図 2 8】



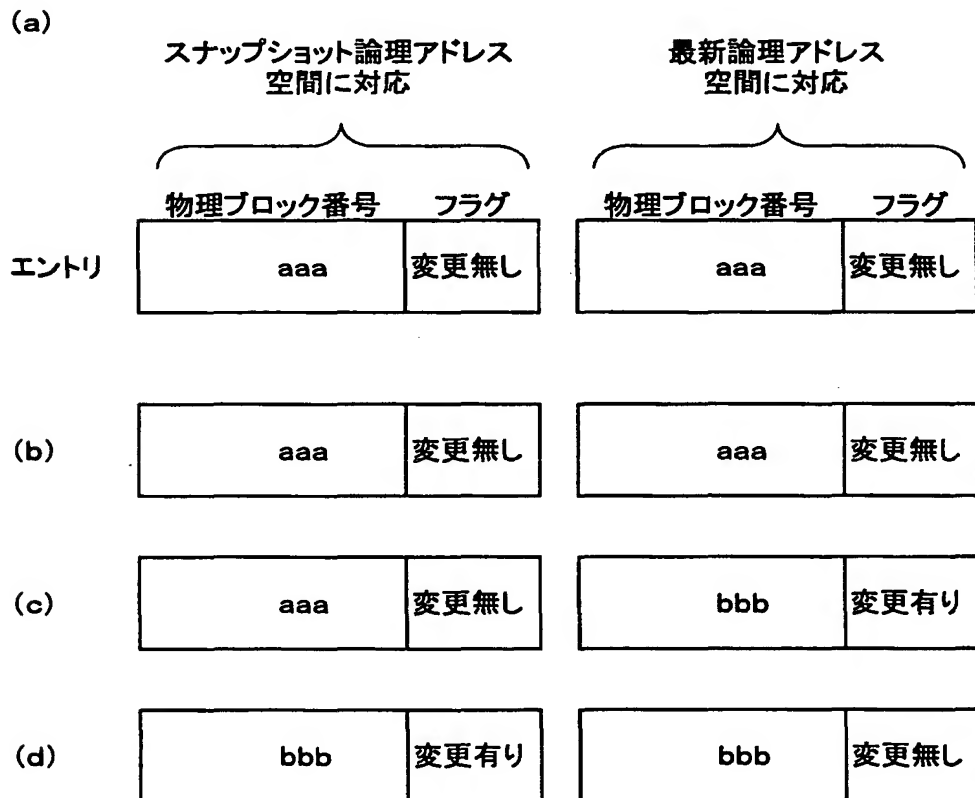
【図 2 9】



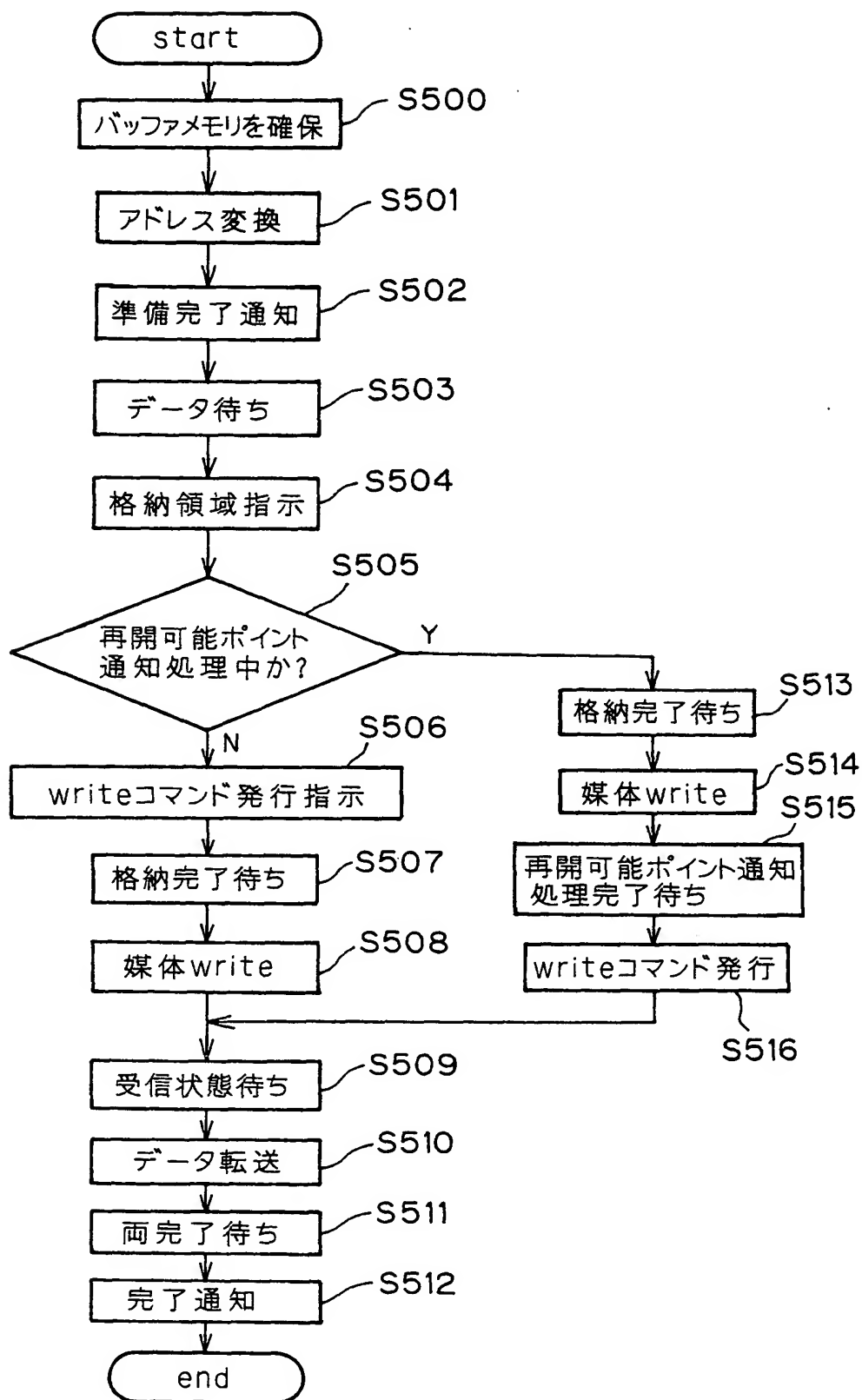
【図 3 0】



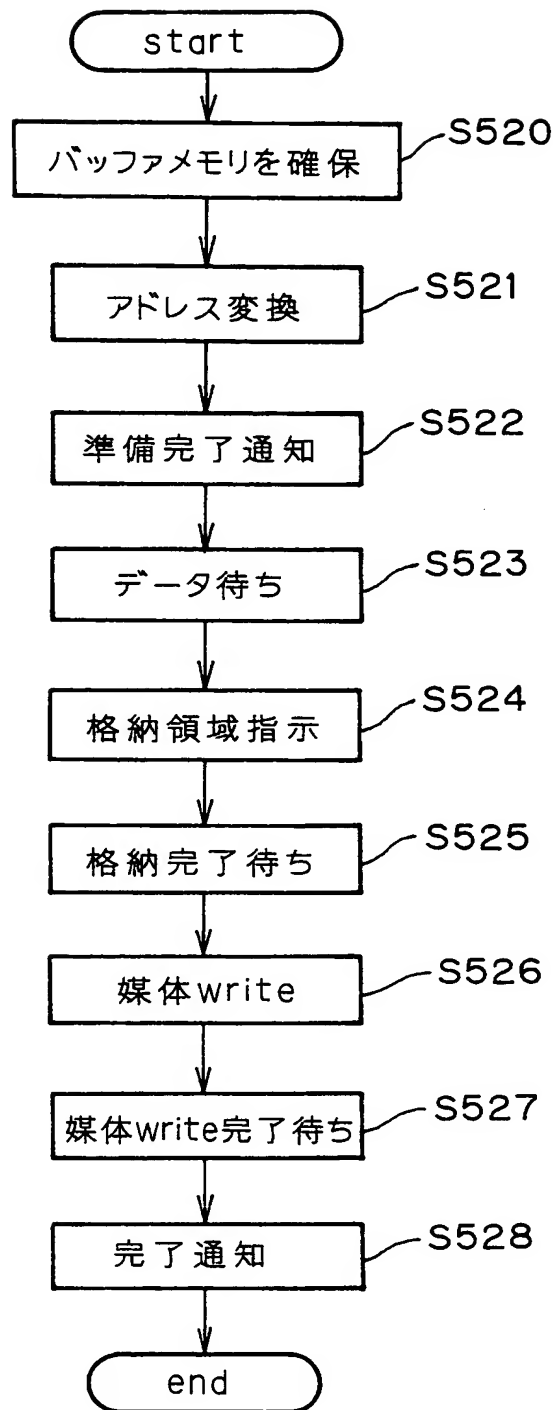
【図 3 1】



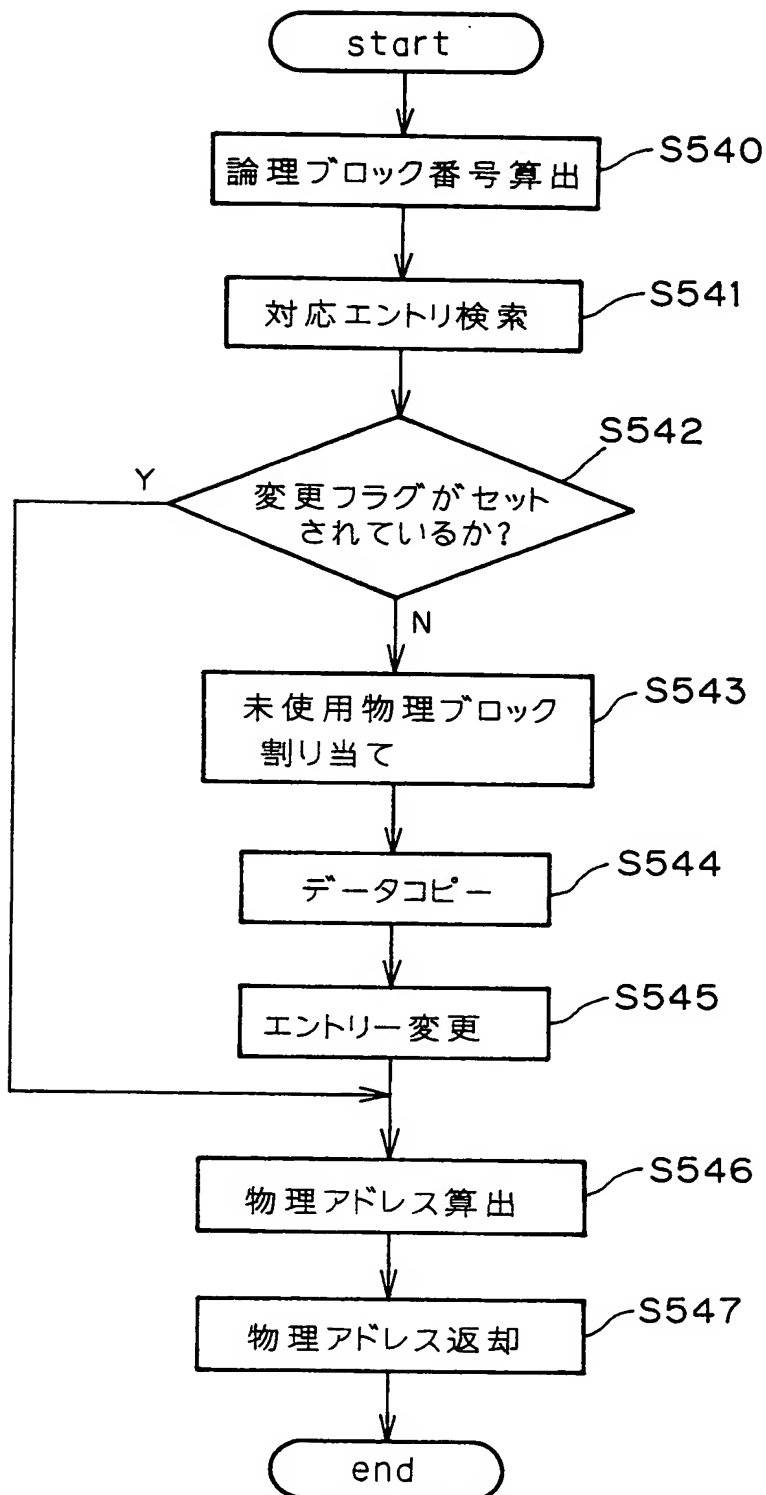
【図 3 2】



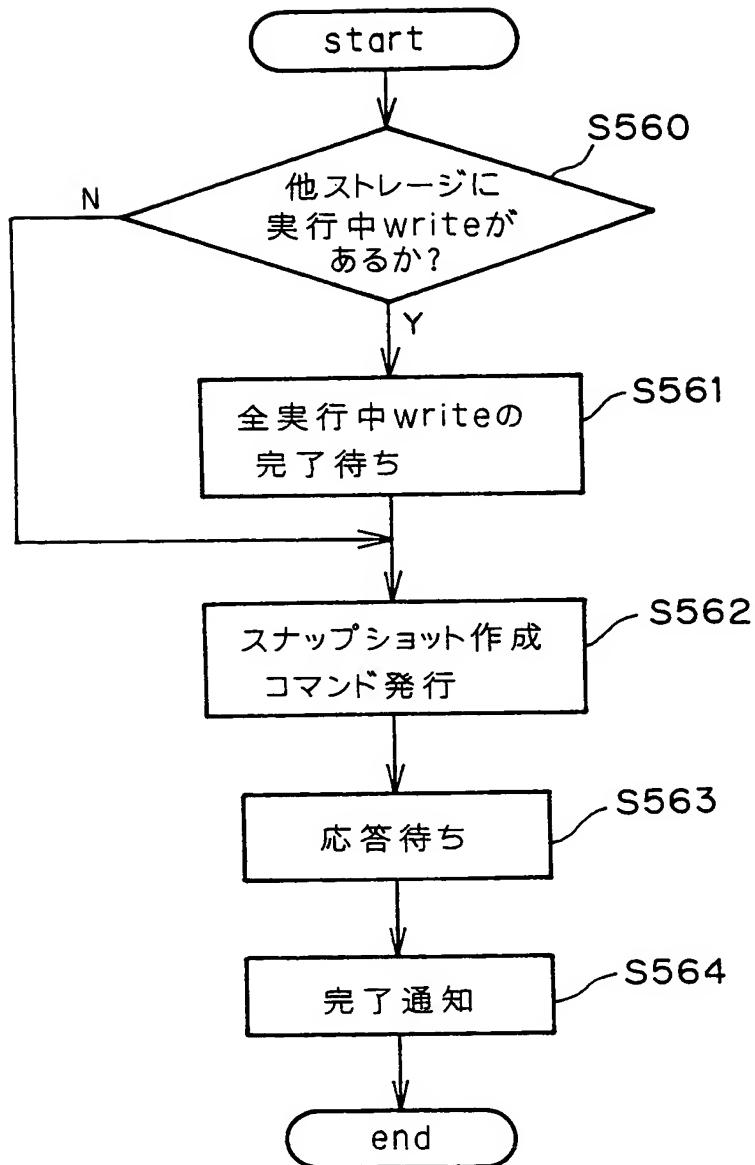
【図 33】



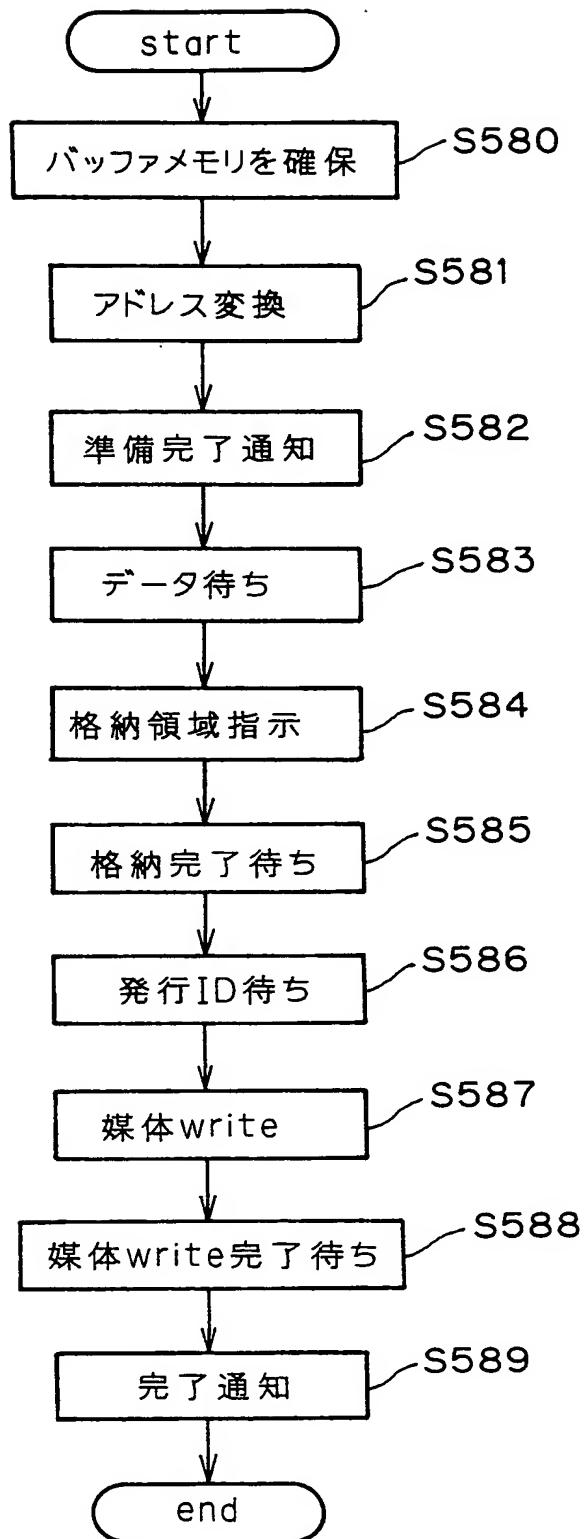
【図 3 4】



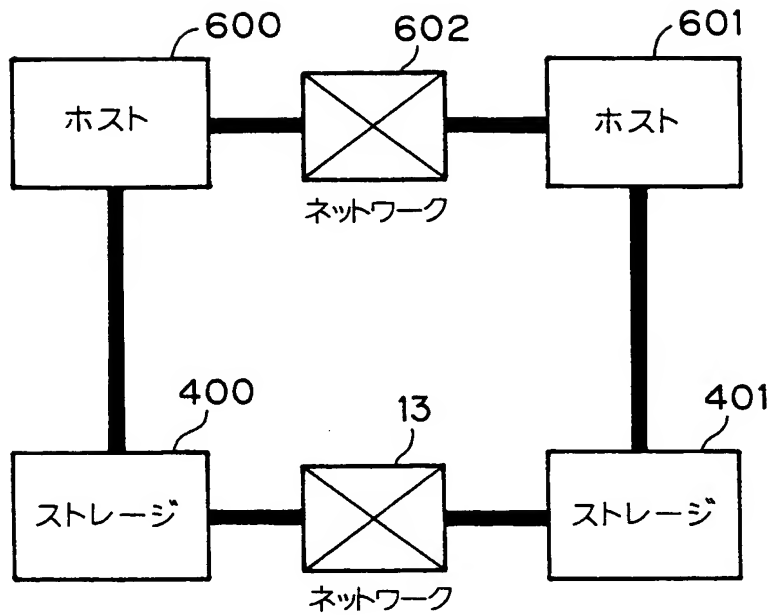
【図 3 5】



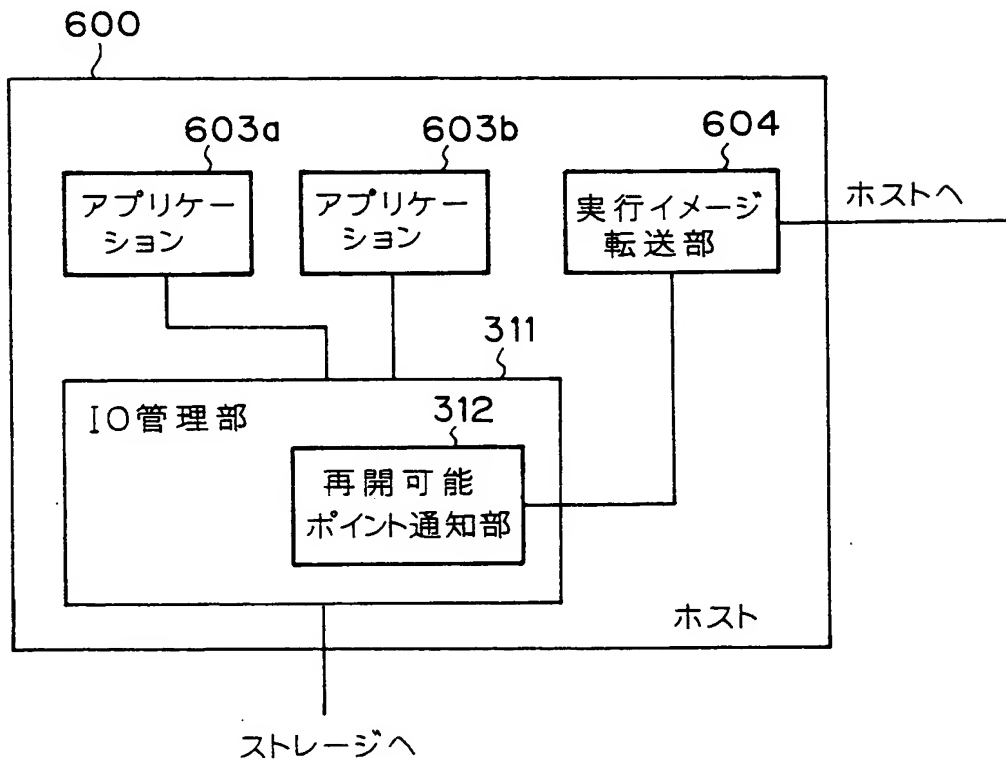
【図 36】



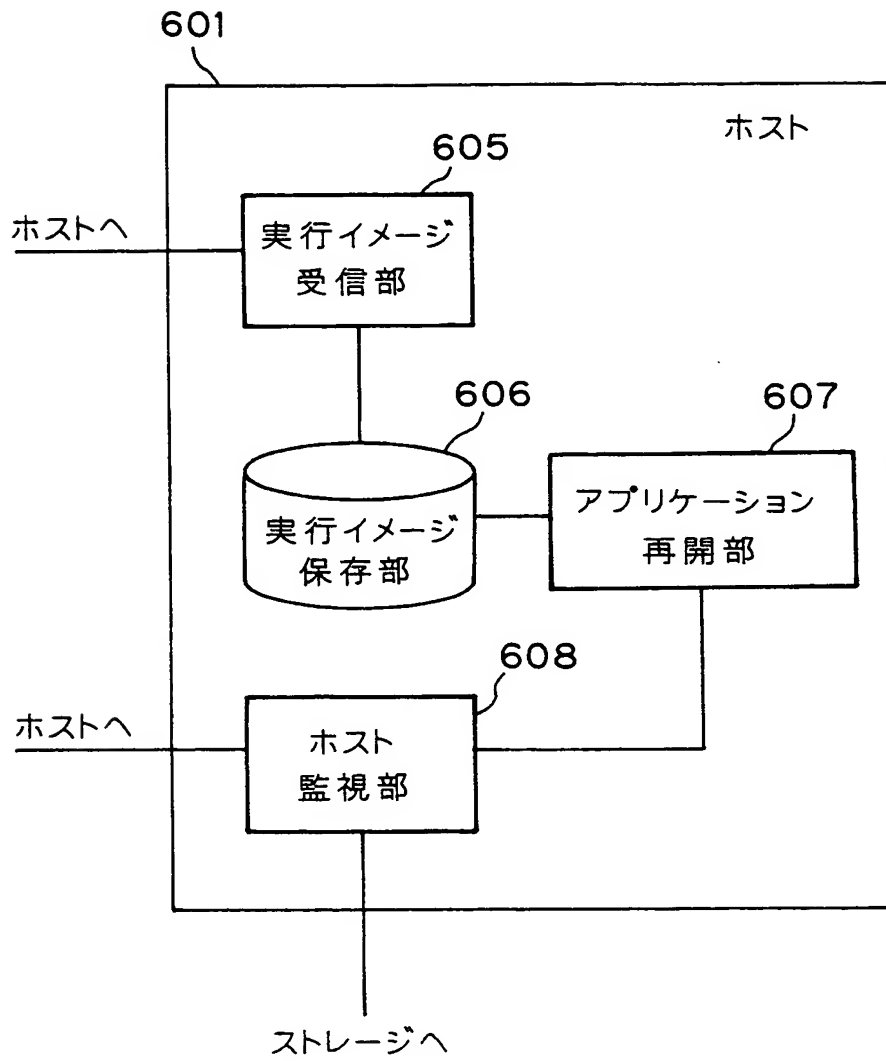
【図 3 7】



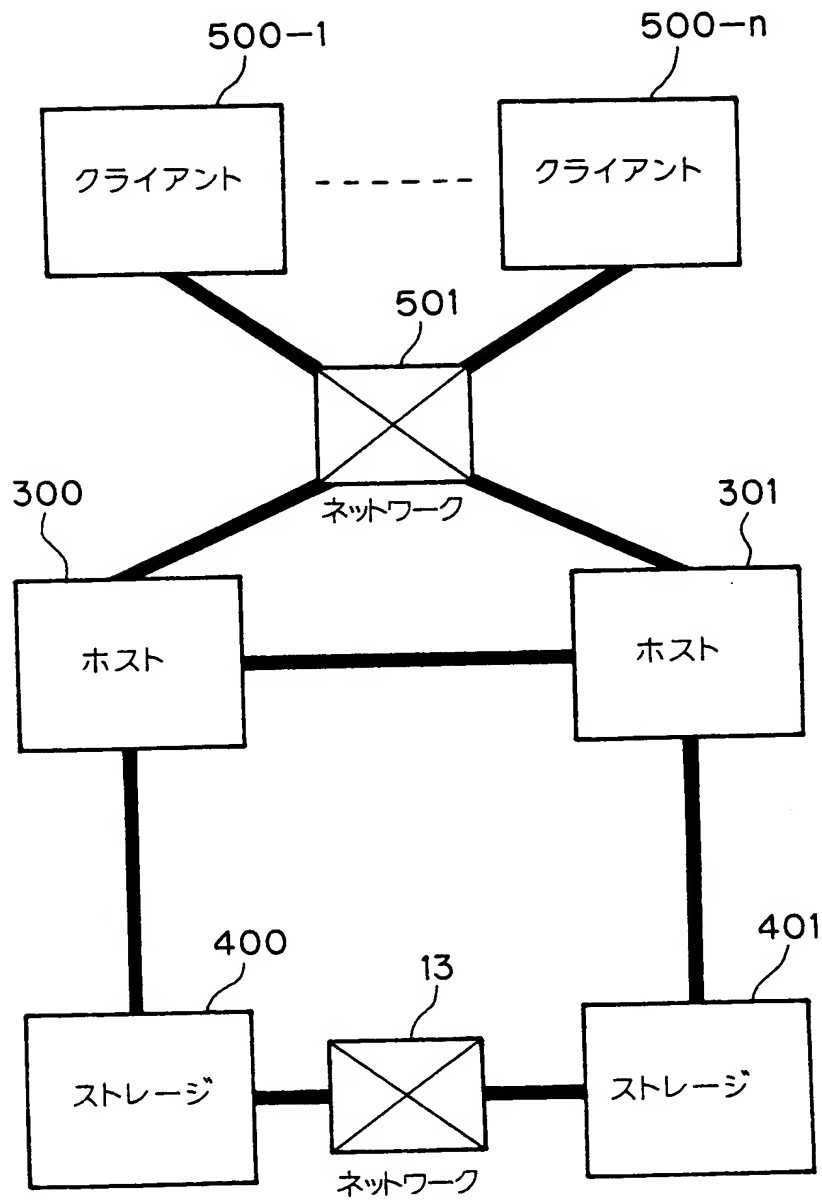
【図 3 8】



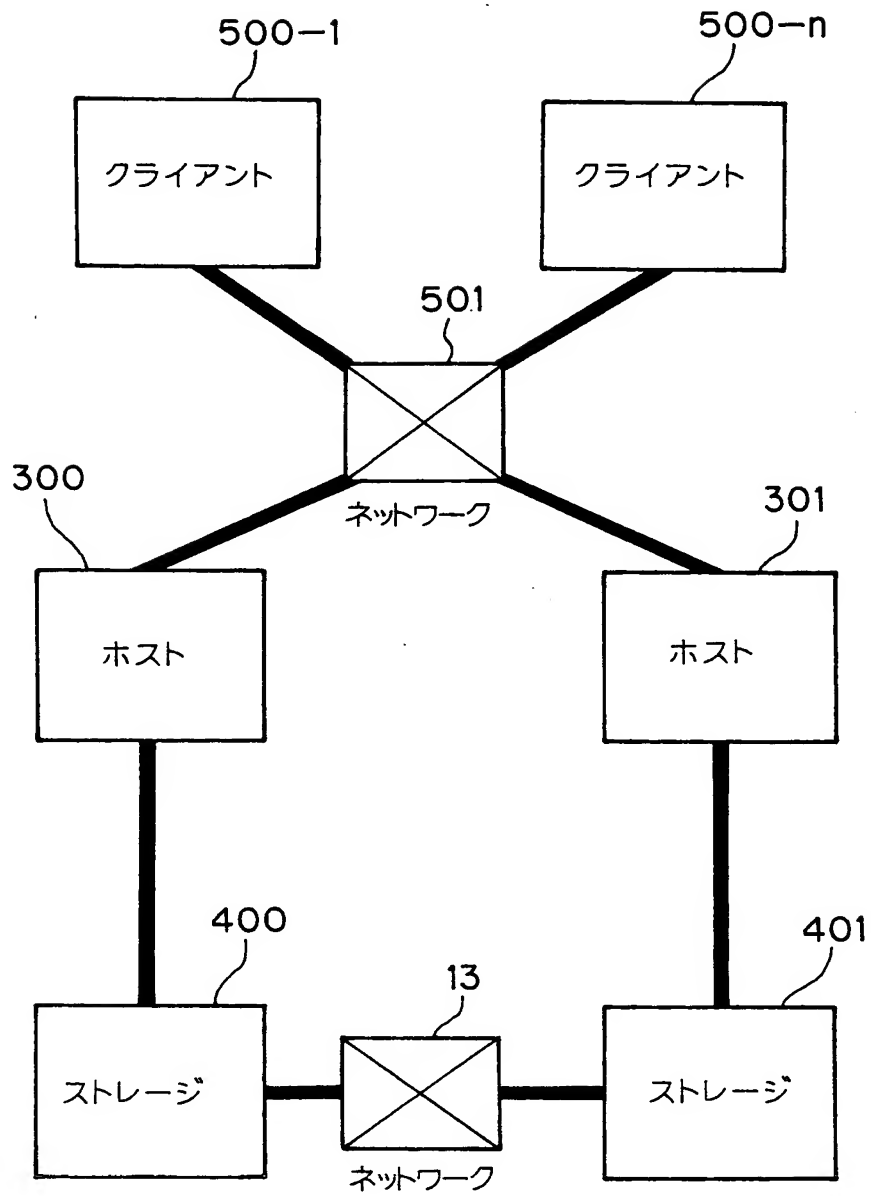
【図 39】



【図 4 0】



【図 4 1】



【書類名】 要約書

【要約】

【課題】 待機系のシステムにデータを送信して耐障害性を確保する際、データの送信に伴う正常系のシステムの処理の遅れを防止する。

【解決手段】 中継装置 1 5 は、正常系のホスト（ホストコンピュータ） 1 0 おおびストレージ 1 1 の設置場所において地震等の災害が発生したときに、災害の影響が波及すると想定される範囲の外に設置される。さらに、ストレージ 1 1 とストレージ 1 2 とが直接データを転送する場合のデータ転送時間よりも、ストレージ 1 1 と中継装置 1 5 との間のデータ転送時間が短くなるような位置に設置される。中継装置 1 5 は、ストレージ 1 1 からデータを受信した場合、ストレージ 1 2 に対してそのデータの送信を完了させる前に、ストレージ 1 1 に対してデータ受信完了を通知する。

【選択図】 図 1

出 願 人 履 歴 情 報

識別番号 [000004237]

1. 変更年月日 1990年 8月29日
[変更理由] 新規登録
住 所 東京都港区芝五丁目7番1号
氏 名 日本電気株式会社